OLUME 104 PART B NUMBER 14 MARCH 1957

The Proceedings

OF

THE INSTITUTION OF ELECTRICAL ENGINEERS

FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

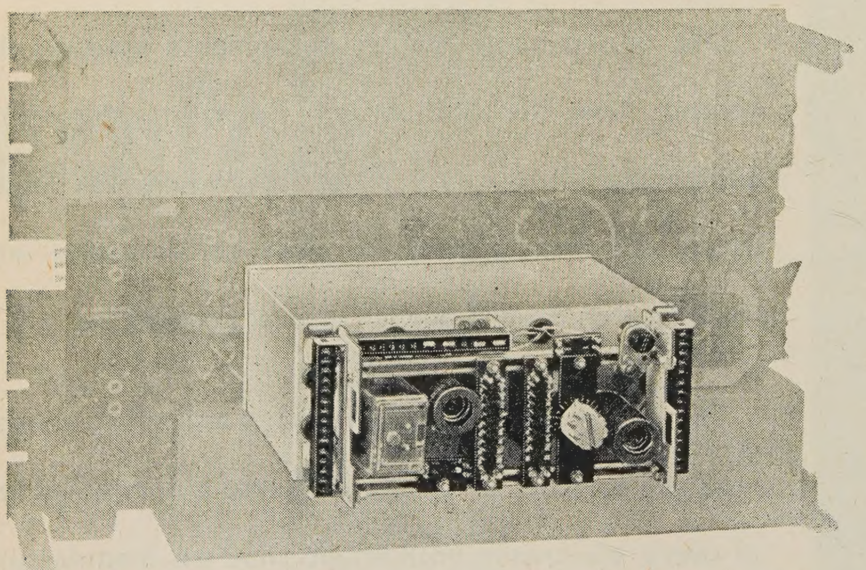
RADIO AND ELECTRONIC ENGINEERING
(INCLUDING COMMUNICATION ENGINEERING)

SAVOY PLACE . LONDON W.C. 2

Price Ten Shillings

G.E.C. **announce**

TRANSMISSION



The new channel unit embodying out-of-band signalling equipment compared with an earlier channel panel without signalling facilities

NEW CHANNEL PANEL

Latest miniaturisation techniques ha been employed in the design of a n channel panel, which includes out- band signalling at 3825 c/s. Compared the channel panel in general use or five years ago—for which the signalli equipment was mounted on a separ panel, and which did not incorpor out-of-band signalling—the new pa occupies only one-sixth of the ra space.

There are many advantages to be gain from the use of out-of-band signallir since the speech and signalling chann are independent. This means that spee and signalling signals can be transmitt simultaneously; consequently the junction relay sets are much simpler than those required with in-band signalling systems. The equipment can easily be converted from ring-down signalling to dialling application—an important feature to those Administrations contemplating trunk-dialling systems in the future.

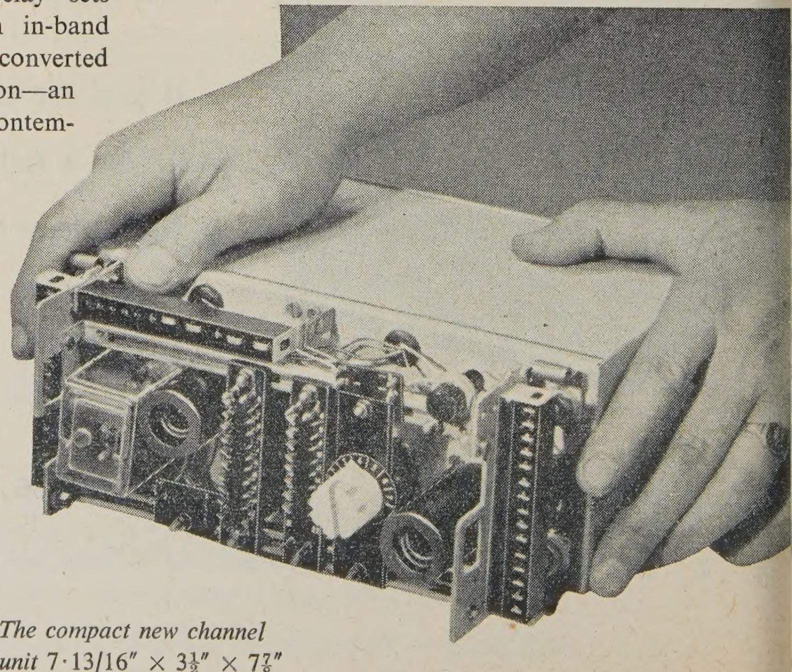
The new channel panel is being incorporated in the following G.E.C. equipment:

OPEN-WIRE EQUIPMENT

A complete terminal for 3-speech circuits plus four duplex telegraph channels, or for 12-speech circuits, can now be mounted on one single-sided rack 9 ft. high × 1 ft. 8½ in. wide. For full information write for standard specifications SPO 1011 and SPO 1025.

BASIC GROUP EQUIPMENT

The complete equipment for three high-quality 12-circuit basic groups can now also be mounted on one single-sided rack 9 ft. high × 1 ft. 8½ in. wide. For full information write for standard specification SPO 3006.



The compact new channel unit 7·13/16″ × 3½″ × 7⅞″

atest developments in

EQUIPMENT

TRANSISTORISED EQUIPMENT

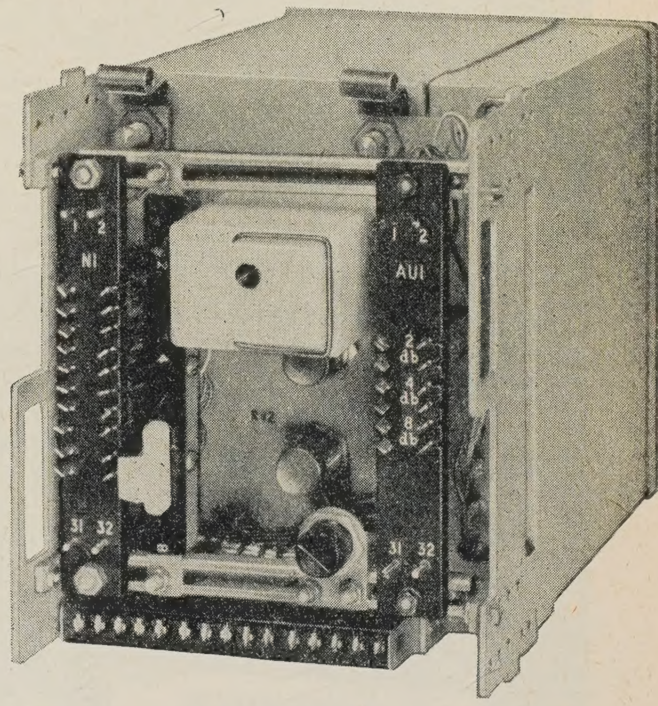
The use of transistors in transmission equipment has many advantages. For example:

The power consumed is extremely low.

The physical size is small.

The heat dissipated is negligible.

The G.E.C. is producing completely transistorised equipment for the following applications:



A complete VF telegraph unit including transmitter and receiver.

JRAL CARRIER SYSTEM

This enables up to ten circuits to be transmitted over e pair of wires, with facilities for terminating one or re circuits at intermediate points. Full information this equipment is given in standard specification O 1030.

DICE-FREQUENCY TELEGRAPH EQUIPMENT

This extremely compact equipment uses transistors oughout, and operates from a 21-volt dc supply. e system employs frequency shift modulation to pro- e 24-duplex telegraph channels operating at a modula- n rate of 50 bands over any four-wire speech circuit t effectively transmits frequencies between 300 c/s d 3400 c/s. A complete terminal is mounted on a single-sided rack 9 ft. high × 1 ft. 8½ in. wide. Full information regarding the equipment is contained in standard specification SPO 1403.

4-WIRE AUDIO AMPLIFIER

This has a maximum gain of 30dB and is capable of delivering a maximum output of +16dBm. The gain frequency distortion over the range 300 c/s to 6 kc/s does not exceed ·5dB relative to 800 c/s.

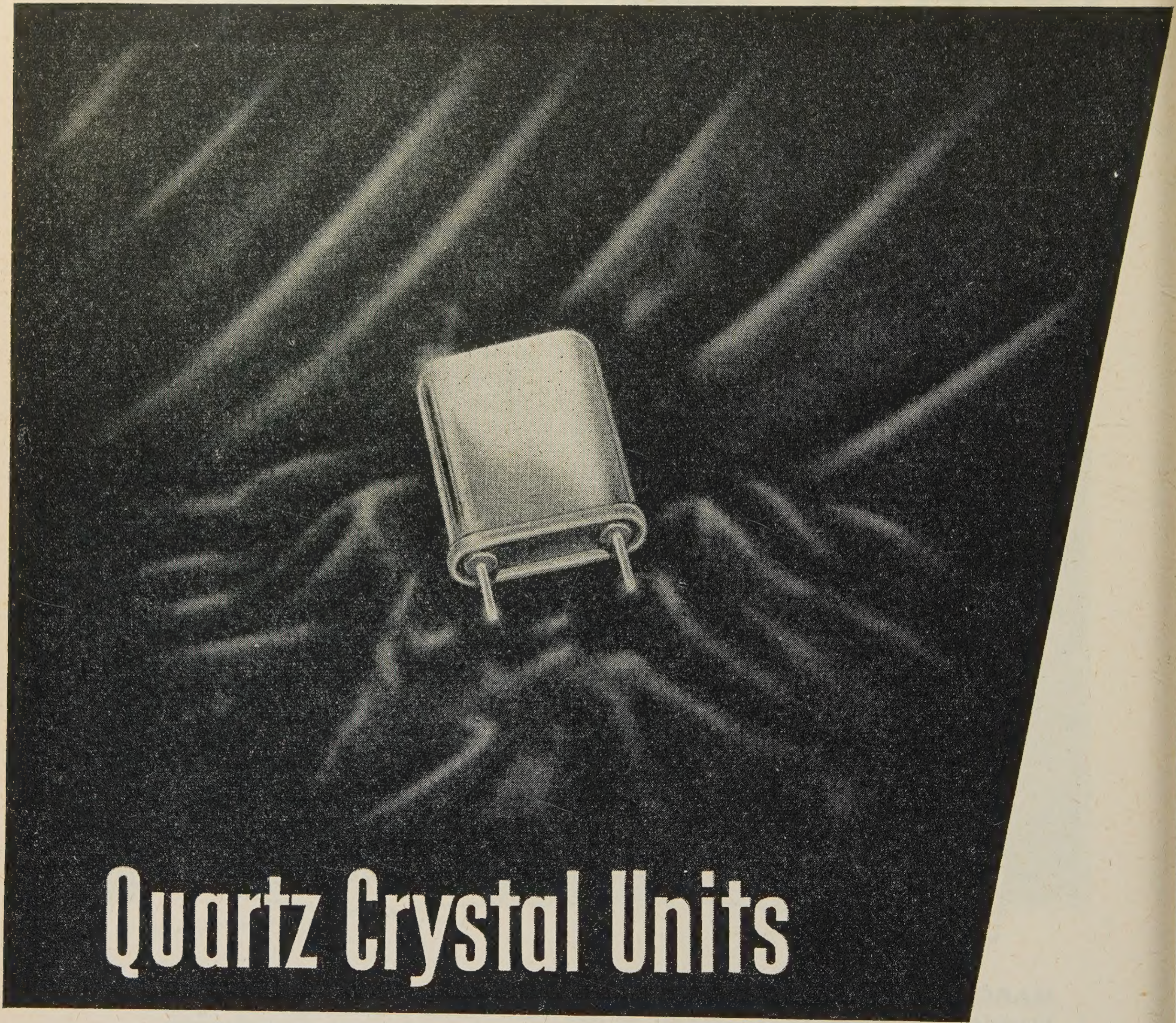
NEGATIVE IMPEDANCE REPEATERS

These are of the shunt and series types for use on loaded 2-wire audio cables.

HE GENERAL ELECTRIC COMPANY LIMITED OF ENGLAND

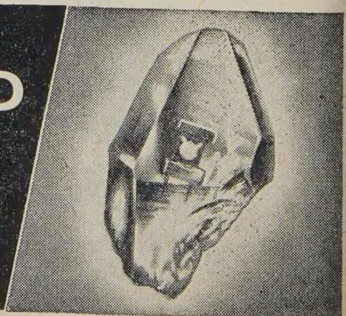
TELEPHONE, RADIO AND TELEVISION WORKS • COVENTRY • ENGLAND

CATHODEON

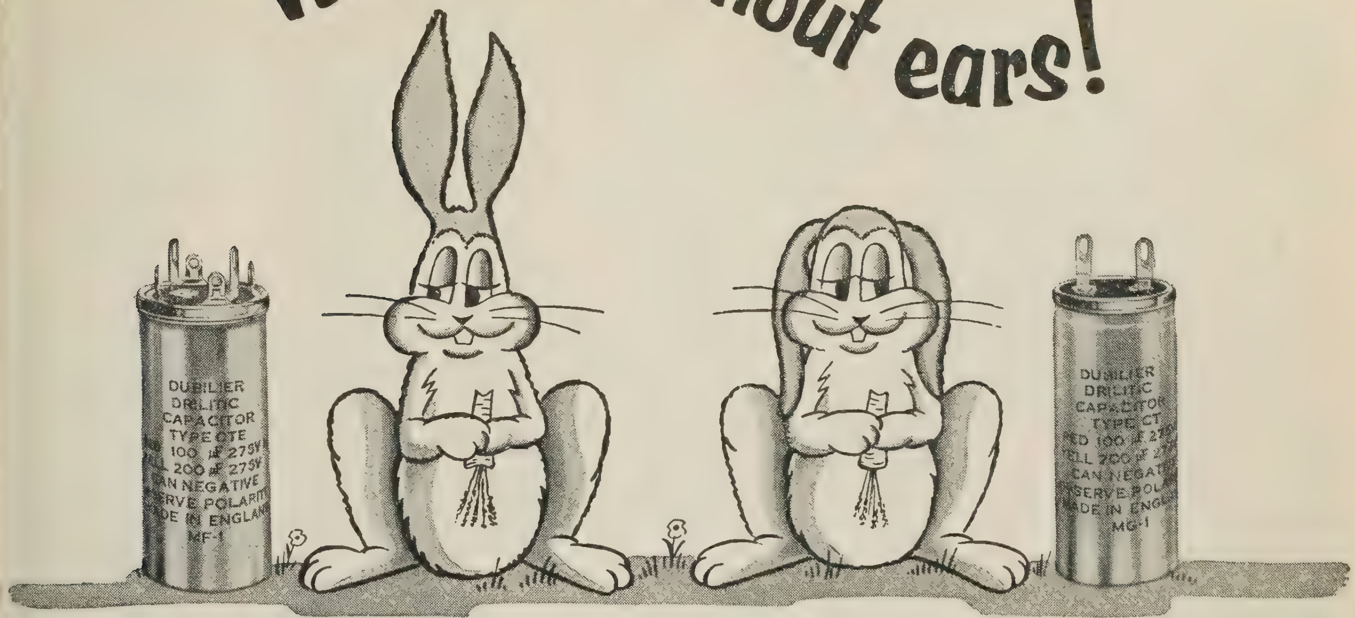


Specialists in HIGH STABILITY CRYSTAL UNITS in the frequency range 2,000-60,000 Kc/s.

CATHODEON CRYSTALS LIMITED
LINTON · CAMBRIDGESHIRE
Telephone : LINTON 223



With or without ears!



Whether you are a radical and like to use modern methods of assembly or whether you are a conservative and prefer the tried and trusted methods, Dubilier can supply you with the capacitors you require.

For example, Dubilier can supply you with electrolytic capacitors for television receivers made either for ear mounting* or clip mounting. In either case they are manufactured with the high ripple current sections required for this purpose.

These capacitors are assembled and sealed in seamless drawn aluminium cans.

**For fixing ear mounting types, only four slots are required in the chassis. The capacitor is dropped into these slots and a slight twist of the ears secures capacitor firmly. Alternatively, a bakelite mounting plate can be supplied for use in those cases where isolation of the capacitor can from chassis is required.*

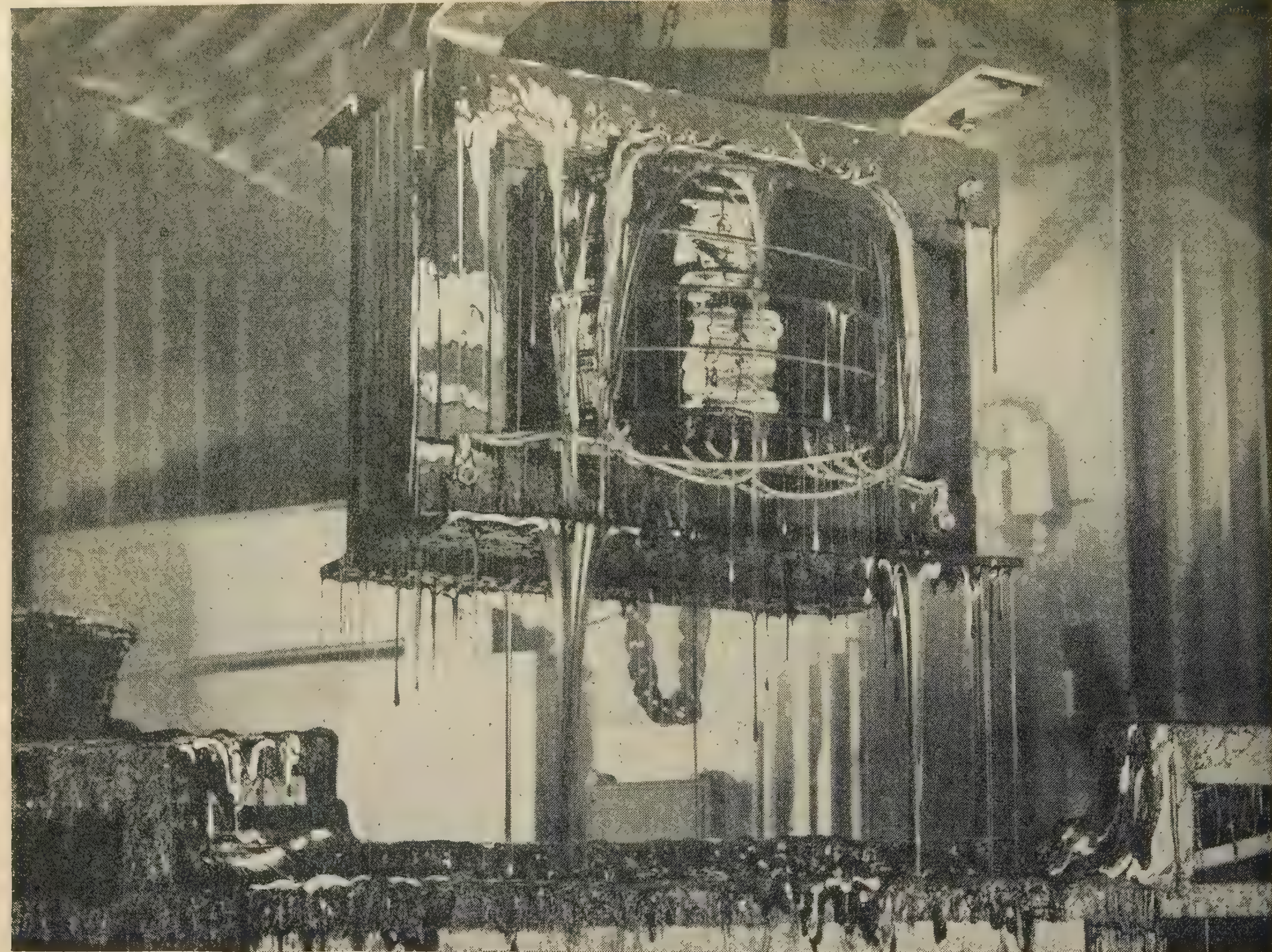
Capacitance (μF)	D.C. Wkg. Voltage	Dimensions	Ripple Current (mA)
100—200	275—275	4″ x $1\frac{3}{8}$″	700—300
50	280	3″ x $1\frac{3}{8}$″	500
100	280	2″ x $1\frac{3}{8}$″	550
50—100	350—280	3″ x $1\frac{3}{8}$″	500—200
50—100	280—280	3″ x $1\frac{3}{8}$″	450—200
200—500	350—350	4″ x 2″	700
64—120	350—350	4″ x $1\frac{3}{8}$″	500
100—200	350—280	4″ x $1\frac{3}{4}$″	900—300
100—200	350—280	4″ x $1\frac{3}{8}$″	700
60—100	350—350	4″ x $1\frac{3}{8}$″	500—200
60—250	350—350	4″ x $1\frac{3}{4}$″	700—400
100—100	350—350	4″ x $1\frac{3}{8}$″	550—200
100—200	350—350	4″ x $1\frac{3}{4}$″	900—300

DUBILIER

Dubilier Condenser Co. (1925) Ltd., Ducon Works, Victoria Road, North Acton, W.3

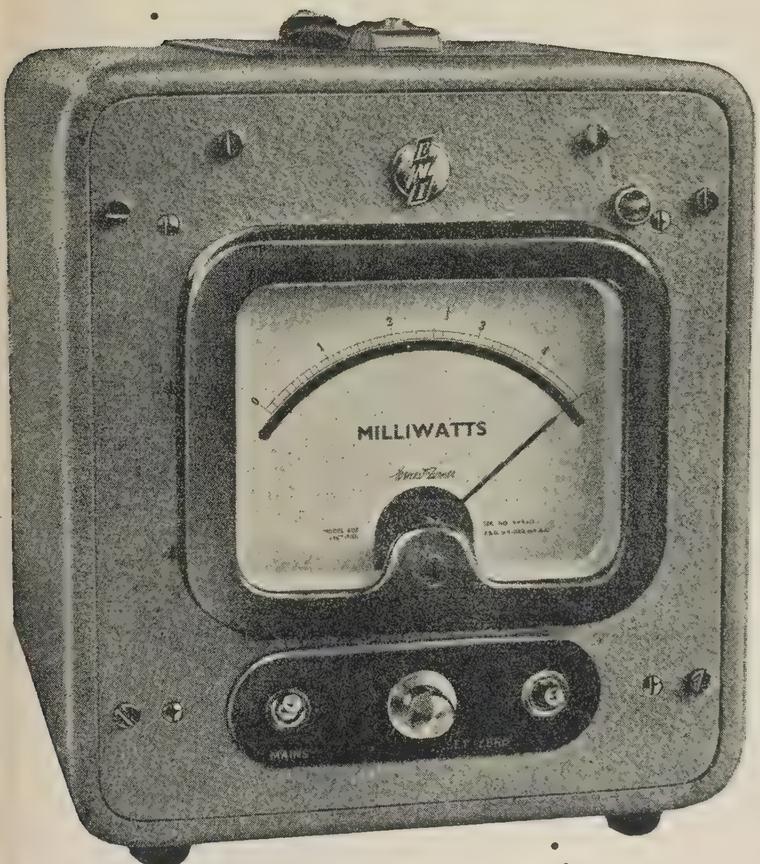
Telephone: ACOrn 2241 Telegrams: Hivoltcon Wesphone, London

DN165B



E·M·I

NEW PRECISION MICROWAVE WATTMETER TYPE 1



with thermistor at constant match during readings

OUTSTANDING FEATURES:

- Accuracy 4%
- Direct Reading
- Constant Microwave Match
- No Thermal Transients to cause zero shift
- Extremely simple to operate

TECHNICAL SPECIFICATION:

POWER RANGE: 0-4mW in steps of 100μW but can be used with a directive feed to measure greater power.

FREQUENCY BANDS: The mount readily available is in W.G.16 and is matched at 9.5 kMc/s. Bandwidth $\pm$250 Mc/s

ACCURACY: Within 4% of ideal.

V.S.W.R.: Better than 0.95 under ideal conditions.

PHYSICAL DIMENSIONS: $10'' \times 10'' \times 10''$ (25cm. $\times$ 25cm. $\times$ 25cm.)

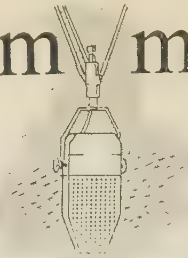
OTHER E.M.I. MICROWAVE EQUIPMENT:

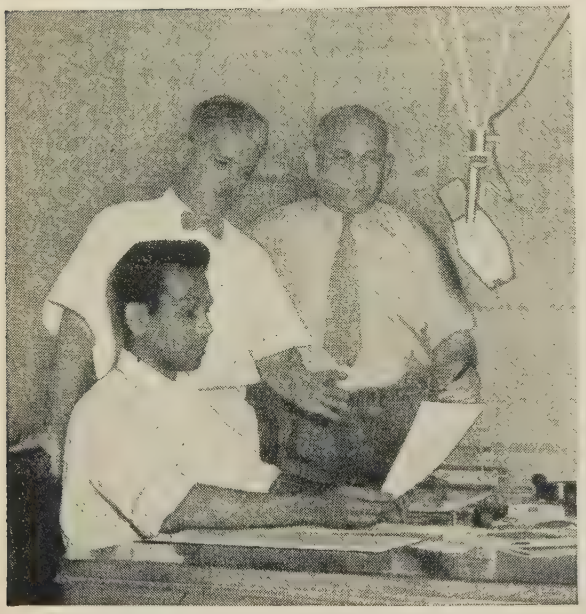
- *Multi-range Cavities*
- *Variable Diameter Cavities*
- *Klystrons & Magnetrons*
- *Stabilised Power Supplies*
- *Selective Amplifiers*

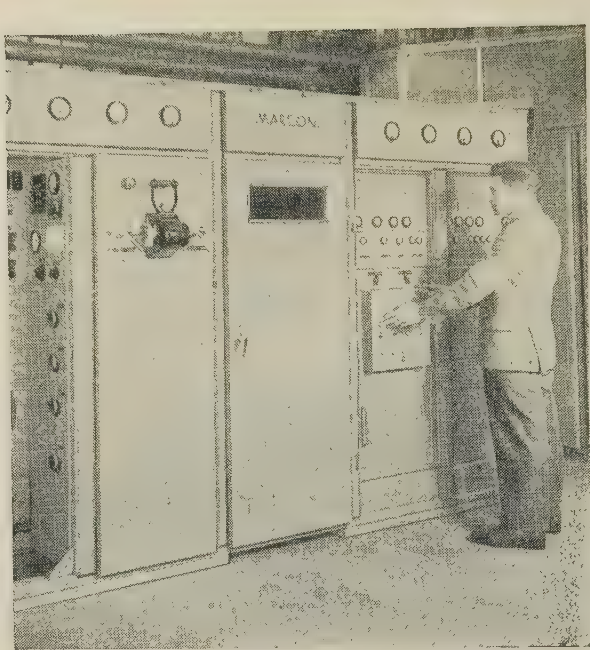
E.M.I. ELECTRONICS LTD.

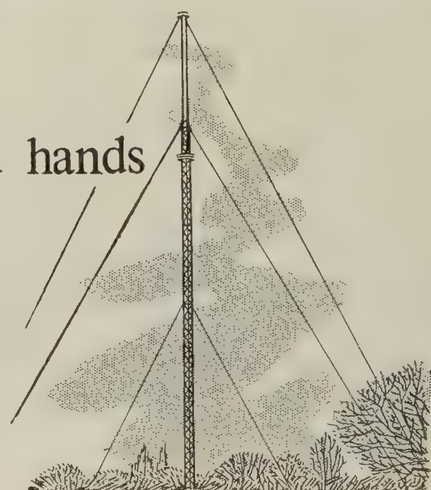
INSTRUMENT DIVISION · HAYES · MIDDLESEX · Telephone: SOUthall 2468 Ext. 1071

EE87

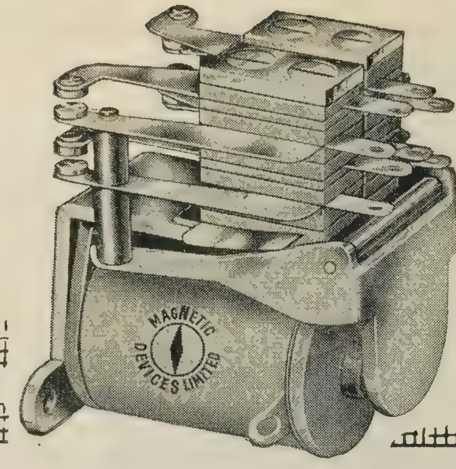




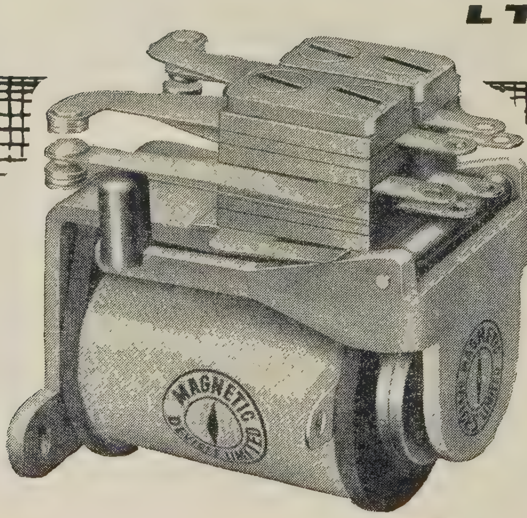




When it's RELAYS contact



Magnetic Devices LTD.



Upper Illustration: Series 596. D.C. operated, Max. V. 140. Contact rating up to 5A continuous. Switching: One to six poles in various combinations. Overall size: 1 7/16″ long by 1 3/32″ wide by 1 25/32″ deep. Coil consumption 0.5 to 3 watts.

Lower Illustration: Series 590. A.C. operated, Max. V. 250. Contact rating up to 5A. continuous. Switching: One to four poles in various combinations. Overall size: 1 7/16″ long by 1 3/32″ wide by 1 25/32″ deep. Coil consumption 2 or 4VA.

Coils are wound for standard voltages up to 250V. A.C and 140V. D.C. Coils can be supplied vacuum impregnated if required.
Please write for illustrated leaflet.

A.I.D. & A.R.B. approved.

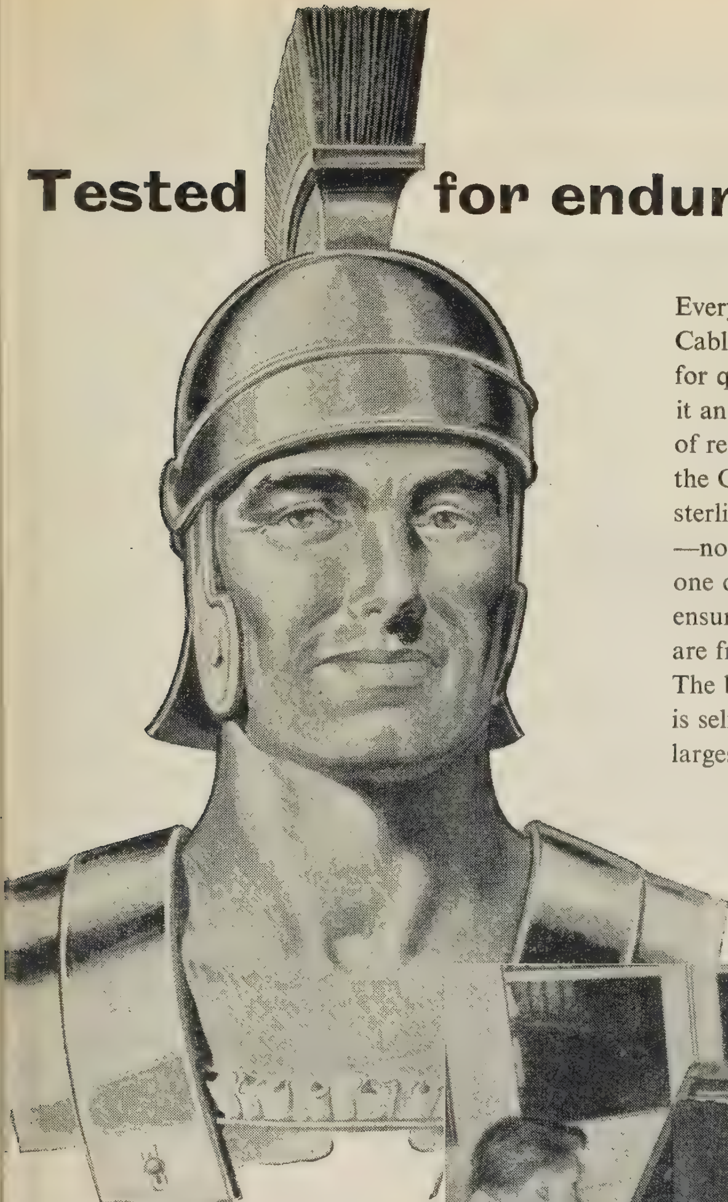
MAGNETIC DEVICES LIMITED

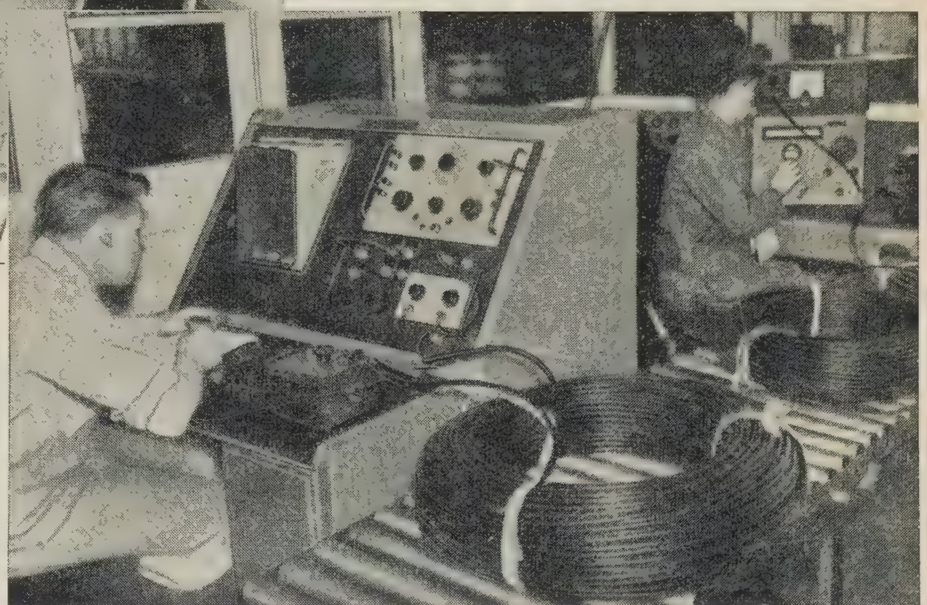
Magnetic Devices LTD.

EXNING ROAD NEWMARKET

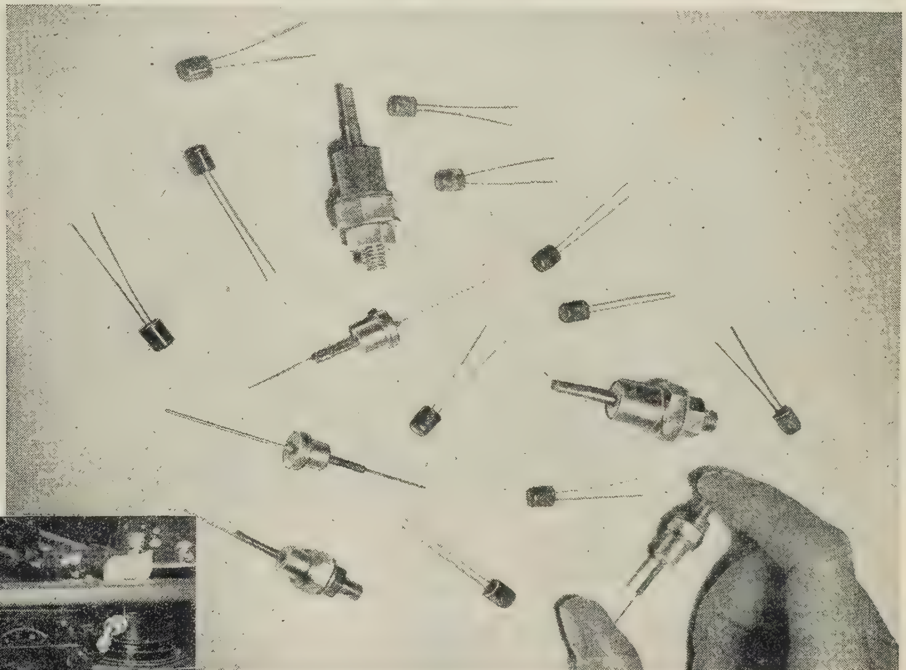
Telephone: Newmarket 3181/2/3.

Telegrams: Magnetic Newmarket.

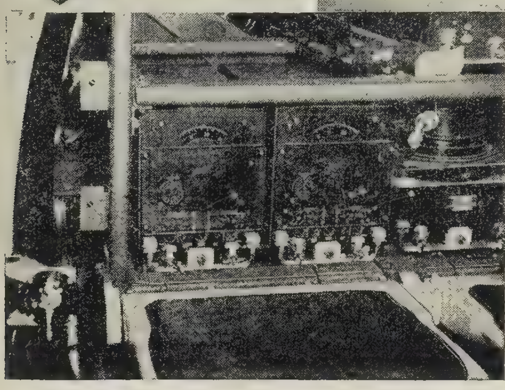




A wide range of applications
FERRANTI SILICON JUNCTION DIODES *and* SILICON POWER RECTIFIERS



Ferranti Silicon Junction Diodes are used in Smiths Flight System.



- LOW REVERSE CURRENT
- HIGH FORWARD SLOPE
- HIGH TEMPERATURE OPERATION
- SMALL PHYSICAL DIMENSIONS
- HIGH MECHANICAL STRENGTH

FERRANTI

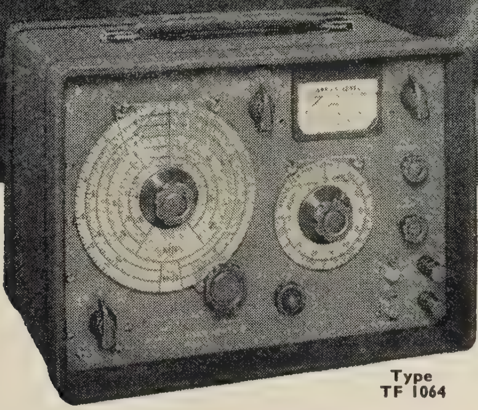
Ferranti Silicon Junction Diodes have been chosen for use in Smiths Flight System not only for their efficient operation, but also for their complete reliability, robust construction, small size and lightness in weight.

Ferranti Silicon Junction Diodes and Silicon Power Rectifiers have many applications in the aircraft, electronic, electrical and general engineering industries including aero engine controls, aircraft power supplies, radar systems, guided missiles, computers, indicating and recording instruments, process control and telephone equipment.

FERRANTI LTD · GEM MILL · CHADDERTON · OLDHAM · LANCS

London Office: KERN HOUSE · 36 KINGSWAY · W.C.2.

FE 174



Type
TF 1064

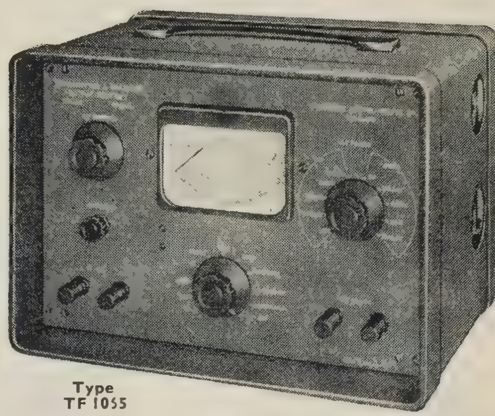
Signal Generator Type TF 1064
Transmitter and Receiver Output Test Set Type TF 1065

The TF 1064 and TF 1065 are complementary instruments. Together, they fulfil all the main testing requirements for mobile transmitter/receiver equipments. Each is completely self-contained and, being light and portable, they are particularly suitable for use in the field.

Signal Generator TF 1064 provides r.f. outputs — f.m. or a.m. — in the ranges 68 to 174 and 450 to 470 Mc/s, i.f. outputs at five spot frequencies, and also an a.f. output.

Test Set TF 1065 comprises an r.f. power meter and deviation indicator for use up to 500 Mc/s, a dual-impedance a.f. power meter, and a multi-range voltammeter.

Each instrument can be supplied separately.



Type
TF 1055

MARCONI
INSTRUMENTS

AM & FM SIGNAL GENERATORS · AUDIO & VIDEO OSCILLATORS
FREQUENCY METERS · VOLTMETERS · POWER METERS
DISTORTION METERS · FIELD STRENGTH METERS
TRANSMISSION MONITORS · DEVIATION METERS
OSCILLOSCOPES, SPECTRUM & RESPONSE ANALYSERS
Q METERS & BRIDGES

MARCONI INSTRUMENTS LTD · ST. ALBANS · HERTFORDSHIRE · TELEPHONE: ST. ALBANS 56161

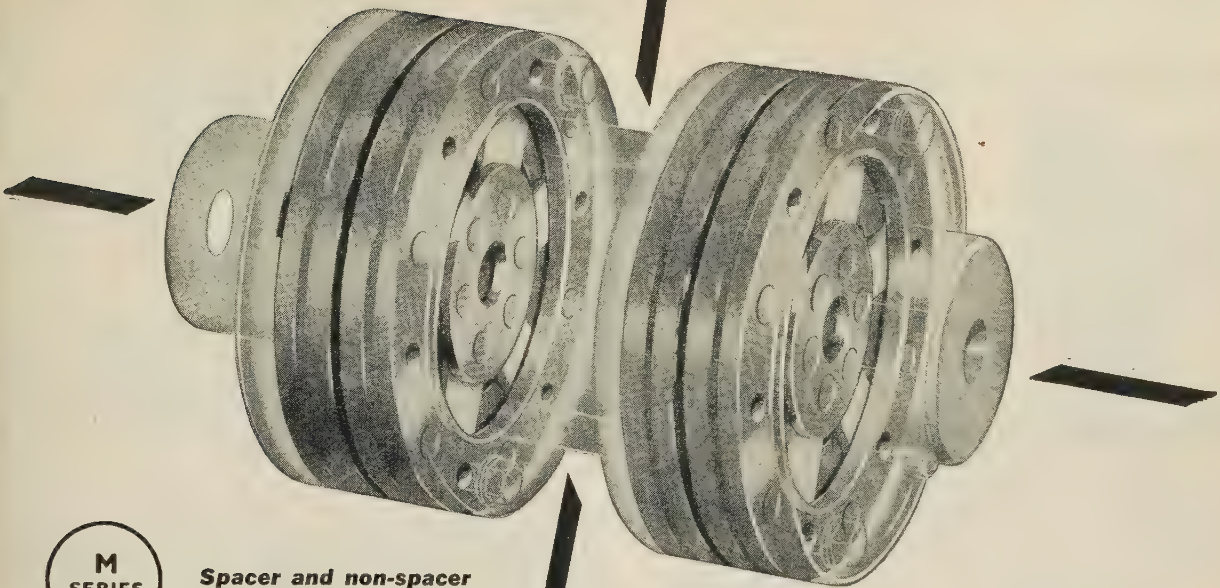
London and the South: Marconi House, Strand, London, W.C.2 Tel: COVent Garden 1234

Midlands: Marconi House, 24 The Parade, Leamington Spa. Tel: 1408 *North:* 30 Albion Street, Kingston-upon-Hull. Tel: Hull Central 16347

WORLD-WIDE REPRESENTATION

FOR CONTINUOUS SERVICE

METASTREAM STANDS OUT



M SERIES — ***Spacer and non-spacer***

LD SERIES — ***Light duty spacer***

SB SERIES — ***Non-spacer for single bearing units***

HS SERIES — ***Cardan shaft units***

. . it protects

Your plant operates under the best conditions ample flexibility being provided, original balance maintained and NO THRUST CREATED.

. . reduces maintenance

Rapid assembly is possible with the compact transmission unit and the spacer coupling gives access to glands or rotors without disturbance to plant. NO LUBRICATION REQUIRED.

. . aids continuous production

All metal construction with flexibility being provided by its strongest element—ripple free stainless steel—ensures long operating life.

Now chosen by refinery and chemical engineers for vital duties throughout the world, METASTREAM is available in over 20 HP/RPM ratings and in four series. Price and delivery are based on the use of the best available materials and economic production methods.

For your next re-equipment or new plant specify:—

METASTREAM *flexible* POWER TRANSMISSION COUPLINGS

the finest of their class in the world

METADUCTS LIMITED

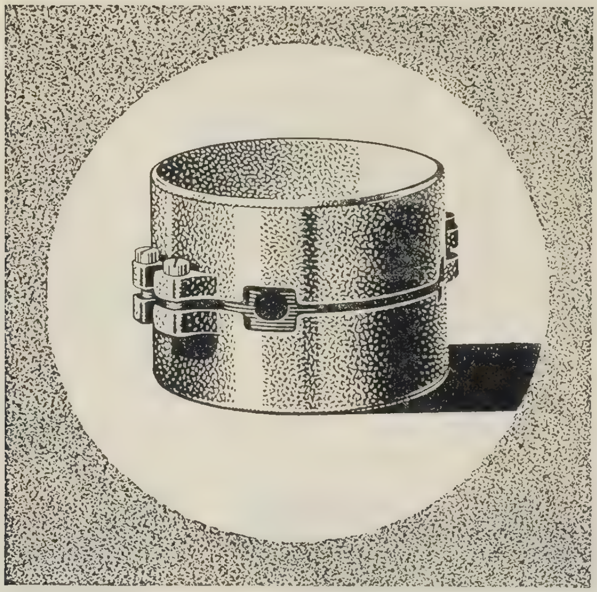
CATHERINE WHEEL ROAD . BRENTFORD . MIDDLESEX . TELEPHONE EALING 3678

A MEMBER OF THE C.M.C. GROUP OF COMPANIES

P4604

NEW GRADE OF FERROXCUBE RESULTS IN

HIGHER EFFICIENCY POT CORES FOR AUDIO FREQUENCIES



Following the development of audio frequency grades of Ferroxcube, Mullard now introduce a series of 30mm pot cores with exceptionally low losses and high effective permeabilities.

Initial permeability μ_0 of 1,400 with a loss of $\frac{\tan\delta}{\mu}$ of 2.5×10^{-6}.

The added advantages of simple construction and light weight make them specially suitable for use in cable splicing and audio frequency output transformers.

Designers requiring full technical details on this new series of high efficiency pot cores are invited to write to the address below.

Mullard

'Ticonal' permanent magnets
Magnadur ceramic magnets
Ferroxcube magnetic cores

MULLARD LTD., COMPONENT DIVISION, MULLARD HOUSE, TORRINGTON PLACE, W.C.1

MC256B

Electronics ...

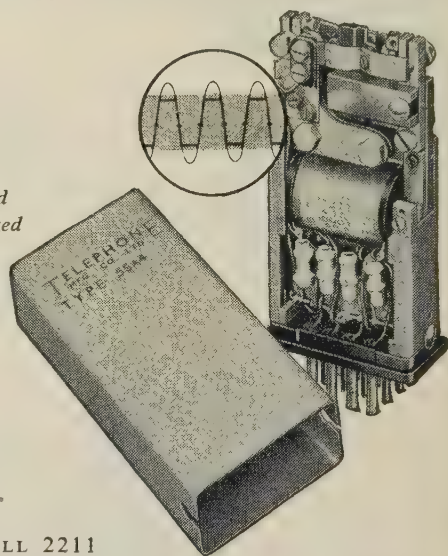
See our exhibits
STAND 72 R.E.C.M.F. EXHIBITION
STAND 201 I.E.A. EXHIBITION

and the Carpenter Polarized Relay

Widespread use has been found for the Carpenter Polarized Relay in Electronic circuits of Industrial and Aircraft equipment.

Its ability to respond to weak, ill-defined, short-duration impulses of varying polarity, and its close operate/release differential has solved many problems of control, amplification, impulse repetition and high-speed switching.

Therefore, if you have a problem which you think could best be solved by a polarized relay—consult us ; our team of Engineers will be only too happy to discuss your requirements with you. But, in any case, write or 'phone for our Brochure F.3516 " Applications of the Carpenter Polarized Relay ".



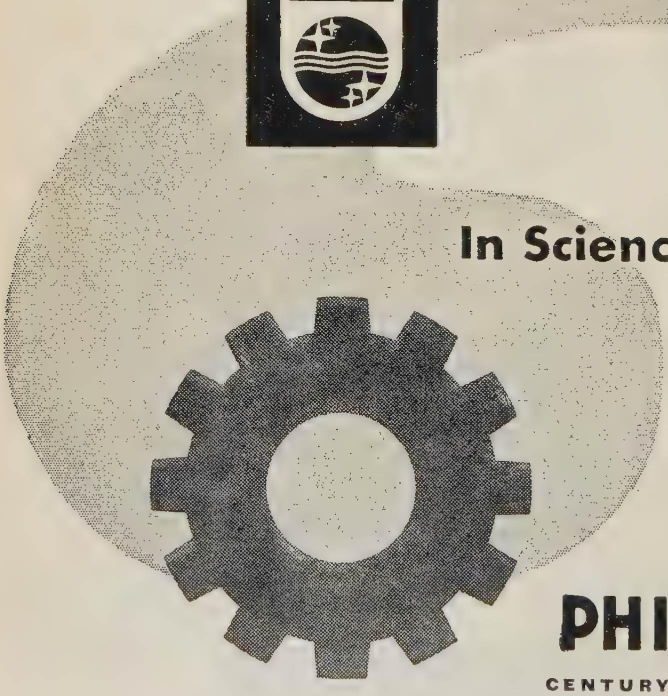
Type 5 Suppressed Carpenter Polarized Relay

Manufactured by the Sole Licensees:

TELEPHONE MANUFACTURING CO. LTD

Contractors to the Government of the British Commonwealth & other Nations

HOLLINGSWORTH WORKS · DULWICH · LONDON S.E.21 TEL. GIPSY HILL 2211



In Science and Industry alike . . .

among technicians, manufacturers and those engaged in the sale of electrical products — as well as among the public at large, the Philips emblem is accepted throughout the World as a symbol of quality and dependability.

PHILIPS ELECTRICAL LTD

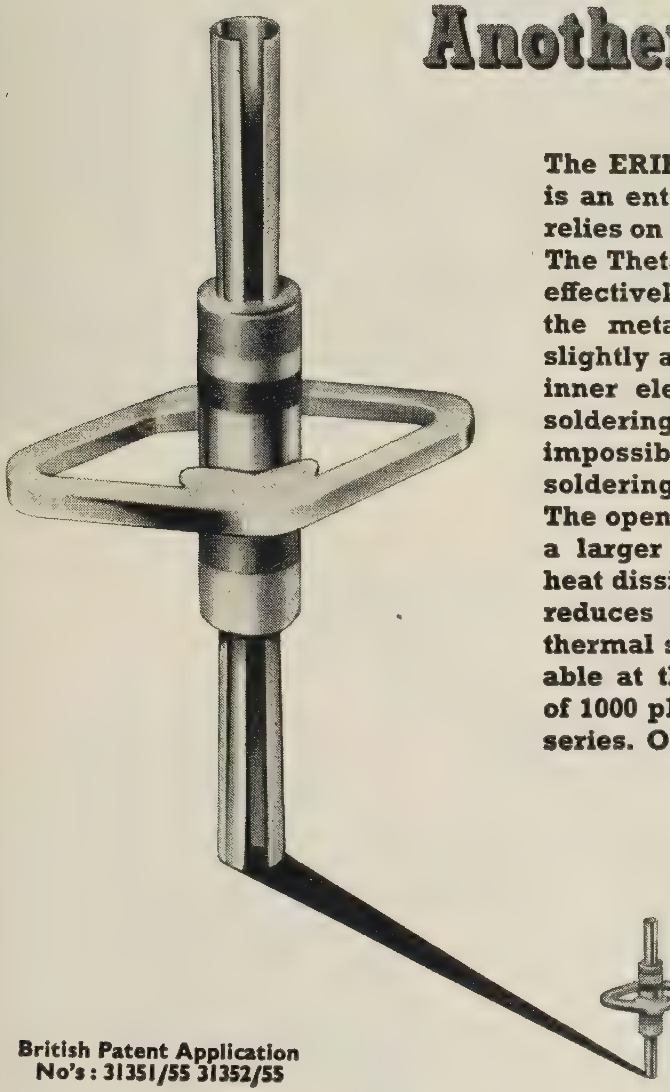
CENTURY HOUSE, SHAFTESBURY AVENUE, LONDON, W.C.2

RADIO & TELEVISION RECEIVERS · RADIOGRAMS & RECORD PLAYERS · GRAMOPHONE RECORDS · TUNGSTEN, FLUORESCENT, BLENDED AND DISCHARGE LAMPS & LIGHTING EQUIPMENT · 'PHILISHAVE' ELECTRIC DRY SHAVERS · 'PHOTOFLUX' FLASHBULBS · HIGH FREQUENCY HEATING GENERATORS · X-RAY EQUIPMENT FOR ALL PURPOSES · ELECTRO-MEDICAL APPARATUS · HEAT THERAPY APPARATUS · ARC & RESISTANCE WELDING PLANT AND ELECTRODES · ELECTRONIC MEASURING INSTRUMENTS · MAGNETIC FILTERS · BATTERY CHARGERS AND RECTIFIERS · SOUND AMPLIFYING INSTALLATIONS · CINEMA PROJECTORS · TAPE RECORDERS

(P23)

THE *Theta* 2000

Another Step Forward



The ERIE Theta Feed-Thru Ceramicon★ is an entirely mechanical assembly and relies on solder only for electrical contact. The Theta mounting, which is sprung on, effectively grips the outer electrode and the metal spill, which is opened out slightly at both ends, expands to grip the inner electrode. Thus, disassembly on soldering to the chassis is absolutely impossible, no matter how prolonged the soldering operation may happen to be.
The open design of the mounting permits a larger hole in the chassis, assists in heat dissipation on soldering, and greatly reduces the possibility of damage by thermal shock. Style 2000, which is available at the moment only in a capacity of 1000 pF – 20% + 50%, is the first of the series. Others will follow in due course.

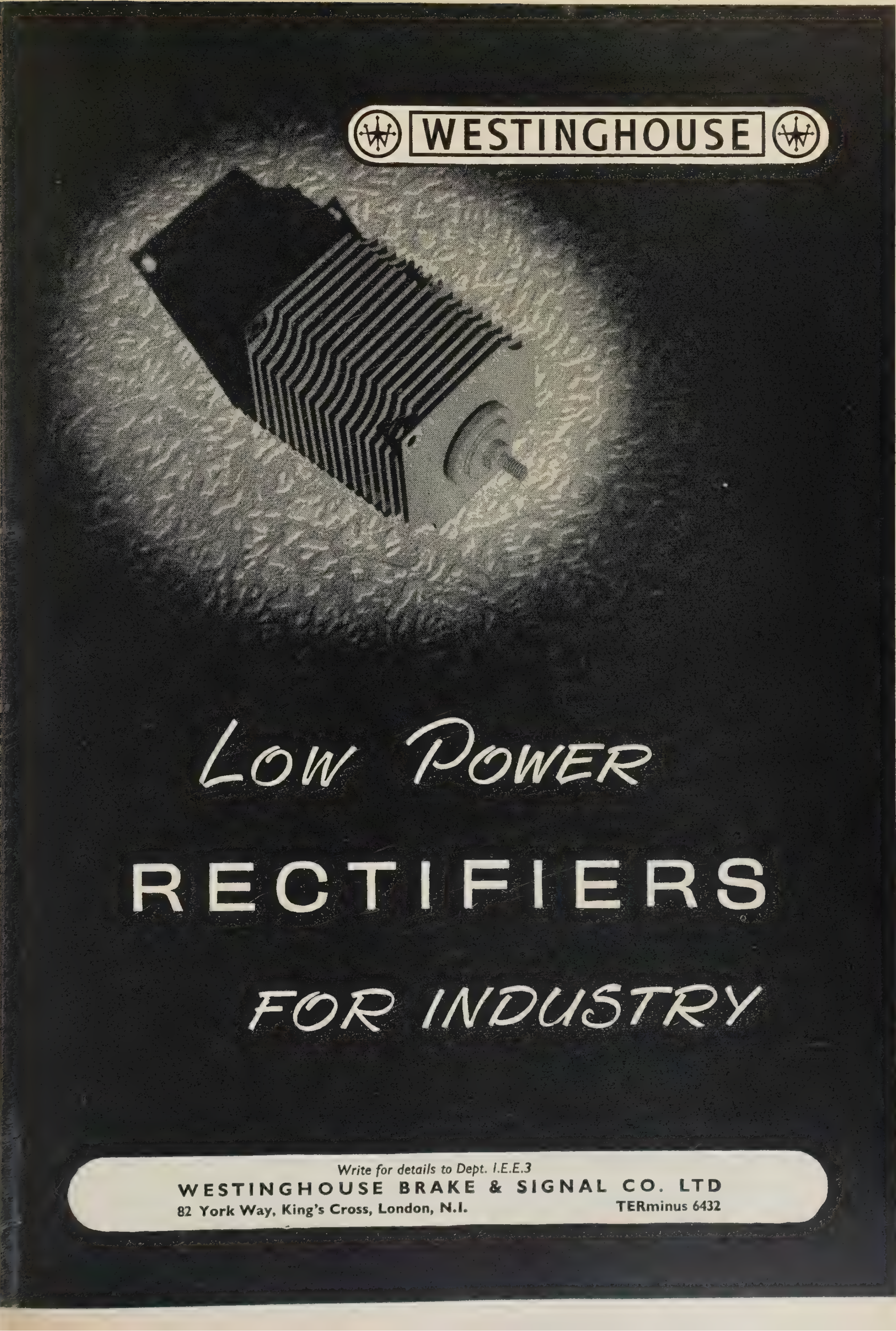
Illustration actual size

British Patent Application No's: 31351/55 31352/55

ERIE★ *Resistor Ltd*

★ Registered Trade Marks

Carlisle Road, The Hyde, London, N.W.9, England. Tel: COLindale 8011. *Factories: London and Great Yarmouth, England; Toronto, Canada; Erie, Pa., and Holly Springs, Miss., U.S.A.*


WESTINGHOUSE
LOW POWER
RECTIFIERS
FOR INDUSTRY
Write for details to Dept. I.E.E.3
WESTINGHOUSE BRAKE & SIGNAL CO. LTD
82 York Way, King's Cross, London, N.I.
TERminus 6432

INDEX TO ADVERTISERS

Firm	page
dcola Products Ltd.	xxiv
ero Research Ltd.	xxxi
rmec Ltd.	iii & xxvi
astinlite Ltd.	ix
utomatic Coil Winder & Electrical Equipment Co. Ltd.	xxviii
utomatic Telephone & Electrical Co. Ltd.	xxxvi
itish Thomson-Houston Co. Ltd.	xxxviii
able Makers' Association	xix
athodeon Crystals Ltd.	viii
nema Television Ltd.	xxv
onnollys (Blackley) Ltd.	xxviii
onovan Electrical Co. Ltd.	xxviii
ubilier Condenser Co. Ltd.	xiii
M.I. Electronics	xv
nglish Electric Valve Co. Ltd.	xxxv
rie Resistor Ltd.	xxxiii
rranti Ltd.	xxi
ortiphone Ltd.	xxxix
X. Fox Ltd.	xxx
eneral Electric Co. Ltd. Telecommunications	iv & v
chard Johnson & Nephew Ltd.	xxiv
ondon Electric Wire Co. & Smiths Ltd.	xxvi
Magnetic Devices Ltd.	xviii
Marconi Instruments Ltd.	xxiii
Marconi Wireless Telegraph Ltd.	vi & vii, xvi
Metaducts Ltd.	xxvii
Metropolitan Plastics Ltd.	xxii
Metropolitan-Vickers Electrical Co. Ltd.	xii
Mullard Magnetic Materials Ltd.	xxix
Mullard Valves Ltd.	i
Newton Bros. (Derby) Ltd.	xxiv
Philips Electrical Ltd.	xxxii
Power Controls Ltd.	xx
Plessey Co. Ltd. "A & A Group"	x
Plessey Co. Ltd. (Swindon)	xvii
Pye Telecommunications Ltd.	xi
Savage Transformers Ltd.	xiv
Solartron Electronic Group	xl
Standard Telephones & Cables Ltd.	ii
Streamline Filters Ltd.	xxx
Telephone Manufacturing Co. Ltd.	xxxii
Westinghouse Brake & Signal Co. Ltd.	xxxvii
Zenith Electric Ltd.	xxx

THE PROCEEDINGS OF
THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 104. PART B. No. 14. MARCH 1957

621.396.11:621.396.933.2

Paper No. 2265 R
Mar. 1957

VERY-LOW-FREQUENCY PROPAGATION AND DIRECTION-FINDING

By F. HORNER, M.Sc., Associate Member.

(*The paper was first received* 27*th July, and in revised form* 16*th October*, 1956.)

SUMMARY

Studies of the propagation of 16 kc/s waves from the Rugby transmitter GBR have been made with the final objective of assessing the polarization errors to be expected in taking bearings on lightning flashes at similar frequencies. Measurements were made of the changes of the apparent bearing of the transmitter caused by changes in the amplitude and polarization of the ionospheric waves. A crossed-loop cathode-ray direction-finder was used for this work.

Polarization errors were largest at distances of about 300 km from the transmitter, where median errors of the order of 10° in daylight and 30° at night were observed. Reasons are advanced for expecting rather smaller errors on a transient signal, such as an atmospheric, but the difference should not be statistically great. Conditions which lead to large errors on a c.w. signal, however, often produce such complex traces with atmospherics that a bearing cannot be read. This is particularly true at night, and common features of night-time observations are missed observations at one or more stations and sets of bearings from which no fix can be derived. A worth-while improvement in fixing accuracy might result from weighting the bearings according to the distance of the flash from each station.

The measurements have yielded further information on the reflecting properties of the ionosphere at 16 kc/s. Variations in bearings taken simultaneously at two stations at similar distances from Rugby, but in different directions, indicated that there were significant differences in the propagation along the two paths. It has been concluded that the polarization of the waves reflected by the ionosphere depends on the azimuthal direction of the propagation path, and this may be the explanation of apparent disagreements between the results of previous workers.

(1) INTRODUCTION

Direction-finding investigations at frequencies near 10 kc/s have been undertaken mainly on account of their application to the location of lightning flashes, and crossed-loop direction-finders have been in regular use for this purpose for many years.[1,2] Polarization errors are one of the main factors which limit the accuracy of these techniques,[3] particularly at night, and the evaluation of the errors is difficult since the true bearing of a lightning flash is rarely known. Information is therefore largely derived from experiments with fixed stations; in the British Isles the most convenient transmitter is GBR, Rugby, on 16 kc/s.

Extensive investigations of the propagation of the waves from GBR have been reported,[4,5] and the probable magnitudes of polarization errors can be estimated from the data. The results of work at different locations are not in complete agreement, however, and it was decided to make direct measurements of polarization errors over a range of distances from the transmitter. The paper therefore deals with two main topics. The first is the measurement of polarization errors on a c.w. signal and a discussion of probable errors on atmospherics. The second is a discussion of the reflecting properties of the ionosphere at 16 kc/s, with particular reference to apparent discrepancies in published results. The information required for these two purposes was derived mainly from a common set of experiments, the techniques for which are described in the next Section. The emphasis throughout was on night-time conditions, in which the largest errors occurred.

(2) OBSERVATION TECHNIQUES

Bearings on GBR were observed on a crossed-loop twin-channel cathode-ray direction-finder (c.r.d.f.) at the Radio Research Station, Slough, at Cambridge and at a number of other locations. The Slough and Cambridge sites were those used in previous investigations. Slough is 108 km from Rugby on bearing 158°, and Cambridge 90 km on bearing 101°. At Cambridge, v.l.f. ionospheric recording equipment of the Cavendish Laboratory[4] was in operation and the work benefited considerably from the data made available for comparison.

In a c.r.d.f., the trace obtained with a vertically polarized field is a diametral line, and the reception of horizontally polarized components of field results in bearing errors and ellipticity of the trace. In much of the work the cathode-ray tube was photographed automatically every minute, and the bearing, ellipticity and amplitude were read from the film. In general, the changes in these parameters from one minute to the next were small, and the films were read at five-minute intervals.

The stability of the direction-finders was such that the twin amplifiers remained aligned over periods of several hours, but as a check, a test signal was applied automatically every thirty minutes and appeared on the film.

On occasion, the traces were read visually at five-minute intervals. Continual monitoring of the bearing in this way enabled the phase of the voltage causing ellipticity of the trace

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

to be determined. Normally there was an ambiguity of 180° in this phase, but by slightly detuning one of the amplifiers and noting whether the ellipticity increased or decreased, this ambiguity could be resolved. The ellipticity could then be accorded a sign, analogous to that of the bearing error. The sign convention was arbitrary, but the changes in sign could be correctly described. Results for a set of measurements in which this technique was used are shown in Fig. 1, which illustrates

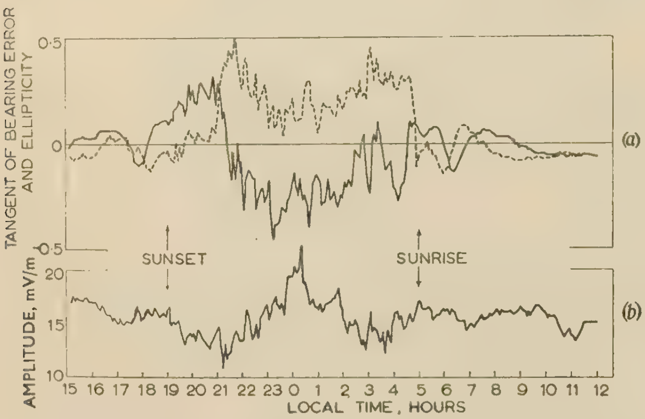


Fig. 1.—Typical direction-finding records at 16 kc/s over a distance of 108 km (29th–30th August, 1955).

(*a*) Bearing data.
——— Tangent of error.
– – – – Ellipticity.
(*b*) Amplitude.

the general features exhibited by all the results at Slough. Daytime errors and ellipse ratios were small and slowly varying. At sunset the gradual increase in the effective height of the reflecting layer introduced phase changes, extending over several hours. These led to a cyclic variation of error and ellipticity, the one being zero when the other reached its peak, and these cyclic changes occurred in reverse order at sunrise. At night the error and ellipticity varied in a random manner and could be large. The relative magnitudes of the error and ellipticity were related to the phase of the ionospheric wave, relative to that of the ground wave, and the average value of this phase at night was found to vary from one night to another.

The bearing data in Fig. 1 are for a good site, and the cyclic variations in the curves are approximately symmetrical about the zero axis. More generally site errors are present and the curves are displaced up or down. These errors introduce difficulties in the derivation of the reflecting properties of the ionosphere, as discussed in a later Section.

The measurements described were all made on c.w. signals, but direct information about atmospherics was also required. This has been obtained from film records taken on a c.r.d.f. and from data kindly provided by the British Meteorological Office from their routine observations. This information was used to confirm the conclusions reached as a result of studying the c.w. data, as described in Section 3.2.

(3) POLARIZATION ERRORS

(3.1) Variation of Errors with Distance from a C.W. Transmitter

Fig. 2 shows a series of typical error curves at distances from 70 to 490 km. The bearings and distances of the sites from the transmitter are shown. At a distance of 70 km, the intensity of the ionospheric wave is much smaller than that of the ground wave and errors are small. As the distance increases, the two waves become more nearly equal at night and the errors increase to a maximum at a distance of about 300 km. Thereafter the

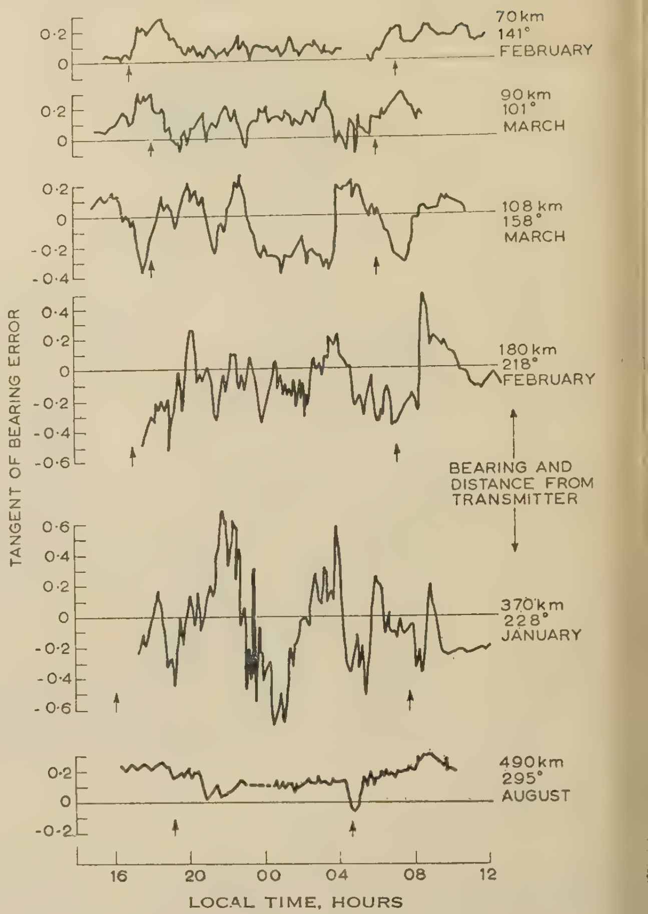


Fig. 2.—Bearing errors at 16 kc/s at various distances.
Times of sunset and sunrise are indicated by arrows.

ionospheric wave begins to predominate, and at 490 km it is about five times the strength of the ground wave, but the angle of incidence becomes progressively less steep and the errors decrease. The first five curves are for winter and the last is for August, but the seasonal variations at night at the longer distances are not large.

The rapid variations in the error and ellipticity at night are caused mainly by variations in the phase of the ionospheric wave, which are apparently random. If, at any instant, this phase could be changed at will, the error could be made to assume a maximum value related to the amplitude of the unwanted (horizontally polarized) component of field, and the ellipticity would then be zero. In direction-finding practice it is usual to compute these maximum errors (called total polarization errors) and to allow for the random phase variations statistically. The total error at any instant is derived from a quadratic combination of the error and ellipticity. The median value of the total error was calculated for each night over an eight-hour period centred on midnight (six hours in summer), and the average of the median values was determined for each receiving site. These averages are plotted in Fig. 3. The Slough (108 km) and Cambridge (90 km) values were based on data for 31 and 7 nights respectively, at various times of the year, those for Penn (86 km) on six nights, mainly in summer, and the rest on only one or two nights, in summer at some stations and winter at others. Seasonal variations, however, are small compared with the variations with distance.

The general shape of the curve, with its maximum at about 300 km, is in agreement with deductions from previous work.

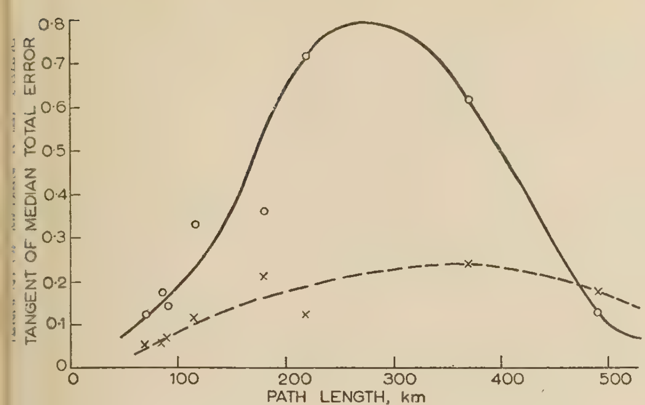


Fig. 3.—Median values of total polarization errors.
——— Night.
- - - - Day.

A similar analysis of errors in daylight leads to the broken curve of Fig. 3. Considerable difficulty was found in allowing for site errors, and the curve must be regarded as indicating only approximate values.

The night-time results at the shorter distances were derived from a large number of measurements, and in later Sections these will be discussed in more detail and compared with previous results.

(3.2) Errors in Bearings of Atmospherics

The magnitudes of polarization errors with c.w. signals having been established, it is necessary to consider whether any significant difference is to be expected in taking bearings on a transient such as an atmospheric. Information will be derived first from a study of the characteristics of the atmospherics and then from statistical data on the quality of fixes obtained with an operational four-station network of direction-finders.

In the first instance it will be assumed that the atmospheric originates at a distance of about 300 km, at which range the largest errors may be expected. The first signal received is the ground wave, followed 150 microsec later by the first ionospheric wave and then by many more echoes. Since the ground wave is vertically polarized, the first $1\frac{1}{2}$ cycles at least (at 10 kc/s) should indicate a bearing free from polarization error. When the ionospheric waves arrive the relative amplitudes and phases of the vertically and horizontally polarized components vary in a complex manner. Successive cycles on the trace may indicate quite different bearings, but the final decay of the trace usually represents only the ringing of the tuned circuits of the amplifiers and is regular, indicating a constant bearing and ellipticity. The bandwidth of the direction-finders was 300 c/s, corresponding to a time-constant of about 1 millisec.

Three differences between this type of signal and a c.w. signal may influence the bearing errors:

(*a*) With a transient, the initial signal from the ground wave only, referred to above, should indicate a substantially correct bearing.

(*b*) The bearing on a transient is normally read on the maximum signal, and the relative horizontal and vertical fields when this occurs will differ from the steady values assumed with a c.w. signal. The bearings with the two types of signal will therefore be different, but it seems unlikely that there will be a substantial statistical difference in the magnitudes of the errors at short ranges. On occasion the errors will be much smaller with the atmospheric either because the maximum signal occurs before the ionospheric wave attains a large amplitude or because the ionospheric waves produce so large and diffuse a trace that only the initial ground-wave trace can be clearly seen. These conditions will occur only infrequently at short ranges.

(*c*) The trace observed with a c.w. signal is always a well-defined ellipse (or in the limit, a straight line), the bearing of which can be read without difficulty, even if the ellipticity is large. With an atmospheric, however, the existence of a large unwanted component of field usually makes the trace confused, because the polarization is changing continually, and there are many traces on which it would be unreasonable to attempt to read a bearing. This is thought to be the main difference between bearings taken on transient and c.w. signals; in conditions which would lead to large errors on a c.w. signal, many of the atmospherics would be ignored.

Examples illustrating these points are shown in Fig. 4. These oscillograms were taken on a rapidly-moving film, with the effect that the later parts of the traces are slightly displaced to the right.

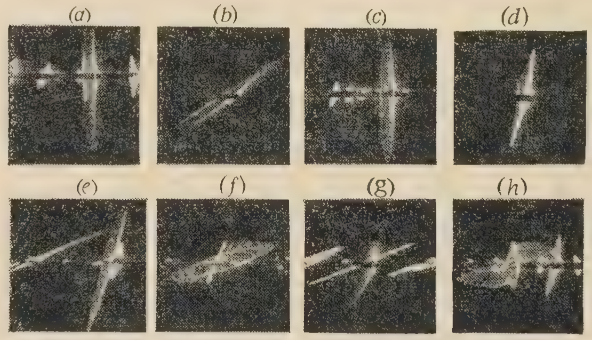


Fig. 4.—Direction-finding observations on atmospherics.

The traces show rotation of the apparent bearings by following angles.

(*a*) Zero (*e*) 58°
(*b*) 4° (*f*) 90°
(*c*) 21° (*g*) 60° } Approximately.
(*d*) 38° (*h*) 90°

In most of the examples the initial part of the trace is clearly visible; in some cases it lasts for three or four cycles, indicating that the atmospheric was from a nearby source. In Fig. 4(*a*) the bearing which would be read on the main part of the atmospheric is the same as on the initial part—there is no swinging due to polarization error, although the trace becomes elliptical. In Figs. 4(*b*), 4(*c*), 4(*d*) and 4(*e*), the trace swings by varying amounts, and there is little doubt that bearings would be taken operationally on the main part of the atmospheric. Figs. 4(*f*) and 4(*g*) are examples in which the main part of the trace shows a large error, but it is possible that, with some traces of this type, the brilliance setting of the cathode-ray tube would enable only the initial part of the trace to be seen, and a correct bearing would then be read. Fig. 4(*h*) shows a more complex trace on which a bearing would probably not have been read.

These examples confirm that large errors exist with atmospherics. From the magnitudes of the errors and the durations of the initial portions of the traces it may be inferred that the atmospherics originated at distances less than 500 km, but their ranges were not, in fact, known.

It is evident that some apparent improvement in accuracy may occur, with atmospherics, by rejection of some of the bearings most affected by polarization error. To investigate the influence of these factors on operational results, an analysis was made of data from the Meteorological Office direction-finding network, obtained over five days in both day and night periods when there were close storms. Errors were assessed for each station by estimating the most probable position of the source, from each set of bearings, and assuming that this was the correct position. This technique tends to minimize the derived errors by distributing them more or less equally between the stations, but should indicate the order of magnitude. The bearings were grouped according to distance, and the r.m.s. error for each distance was determined. Analysis of those sets of bearings from which fixes had been derived showed an increase in the r.m.s. errors from day to night, the maximum value being seven degrees at a distance of about 300 km. The general behaviour was therefore as

expected, but the errors were much smaller than with the c.w. signals. The reason for the smaller errors appeared to lie in the number of sets of bearings which had been rejected, the data for which are shown in Table 1.

Table 1

DETAILS OF FIXES AND REJECTED SETS OF BEARINGS

		Range	
		Less than 500 km	Greater than 1 000 km
Day	No. of fixes	87	136
	No. of rejected sets ..	2	25
Night	No. of fixes	85	107
	No. of rejected sets ..	29	42

Where sets of bearings have not resulted in an acceptable fix, the range is, of course, very approximate, but it is considered that the numbers allocated to the two broad ranges of distance are of the correct order. The Table shows a slight increase, at night, in the proportion of rejected data at large distances and a much greater increase at short distances. Not all the rejections can be attributed to polarization error; some are caused by different observers reading different flashes, when two occur close together in time. This effect is more important at the longer ranges. The sets of bearings for the short distances have been examined, and it has been concluded that at least 18, and probably more, sets at night were rejected as a result of polarization errors.

At ranges greater than 1 000 km the proportion of rejected sets of bearings is higher than at the short ranges, although the polarization errors would be expected to be considerably smaller. One reason is that there is more chance of the different observers reading different flashes when these are small and more frequent. Another is that at 1 000 km the bearings from the four stations are all within a sector of 30°, and relatively small errors can make the estimation of distance difficult.

It is at these longer ranges that there may be a significant difference between polarization errors on atmospherics and c.w. signals. There is considerable delay between the arrival of the low-angle waves which have been reflected only once or twice at the ionosphere, and which give the maximum signal, and the high-angle waves which have been reflected many times. The pick-up on a loop aerial from the horizontally polarized components of the low-angle waves is small, and a bearing taken on an atmospheric when it attains its maximum value is therefore not greatly affected by polarization errors. The bearing on a c.w. signal is taken on the vector sum of all the ionospheric waves, since there is no time discrimination, and is therefore more affected by the pick-up from the horizontally polarized components of the high-angle waves.

From these various theoretical considerations and experimental results it has been concluded that, although there are one or two factors which tend to reduce errors on a transient signal as compared with a c.w. signal, the difference is probably small if all the data are taken into account. On the other hand, since there will be operationally a tendency to read more of the traces with small errors and fewer of those with complex traces, there may be an apparent reduction in errors if only the recorded bearings are studied. It is difficult to assess this effect quantitatively, because it is largely subjective. Perhaps the best use to which the information on polarization errors can be applied is in weighting the various stations according to their distance from the source. For the British network of four stations, if it is assumed, using the results shown in Fig. 3, that large errors are likely to occur at ranges between 150 and 400 km at night, nearly all observations on storms within 400 km of the coast of the British Isles will be subject to large errors at one station at least. In a smaller area they will be subject to errors on two stations, and there are areas in the centre of the network where three or even all four bearings will be subject to large errors. Fortunately, at the shorter distances, large bearing errors do not lead to large positional errors, but it may be possible to obtain a worth-while improvement in accuracy at night by taking into account the probable errors on each station as a function of distance.

There is, no doubt, a similar effect in the day-time observations, but the errors are much smaller and it is doubtful whether a worth-while improvement in accuracy could be obtained by the use of weighting factors.

(3.3) A Note on the Sector Recorder

Many atmospherics direction-finders operate on the principle of the sector recorder.[7,8] The aerial system consists of two parts; one is a loop with a figure-of-eight polar diagram, and the other is a combination of a loop and a vertical rod with a cardioid polar diagram. The maximum of the cardioid is considerably smaller than that of the figure-of-eight and is in line with one of the nulls of that diagram. An atmospheric is recorded only if the output from the cardioid is greater than that from the figure-of-eight, i.e. if it arrives from a direction in the region of a particular null. The whole system rotates at uniform speed, and the atmospherics are recorded on a drum which rotates in synchronism.

The system is essentially a loop direction-finder and the bearing errors are similar to those encountered with the c.r.d.f. On complex atmospherics, however, some differences may be expected because the displays are of different forms. When the sector recorder receives waves with a substantial horizontally polarized component, both the figure-of-eight and cardioid patterns are distorted. The figure-of-eight minima are not nulls and the signal at minimum may be greater than that from the cardioid maximum, which is affected to a smaller extent by the horizontally polarized field. In these circumstances the atmospheric would not be recorded, which is somewhat analogous to ignoring the complex traces on the c.r.d.f. However, c.r.d.f. traces which are well-defined ellipses will be read, and will often yield bearings with small errors, since, if only the phase variations are considered, large ellipticity tends to be associated with small errors (see Fig. 1). These atmospherics may not register on the sector recorder for the reasons given above, because large ellipticity on the c.r.d.f. corresponds to flat minima on the figure-of-eight pattern of the sector recorder.

There may be a tendency, therefore, for the errors to be slightly less, statistically, in a c.r.d.f. than in a sector recorder, but the difference will depend on the way in which the records from the instruments are used.

(4) THE REFLECTING PROPERTIES OF THE IONOSPHERE AT 16 kc/s

(4.1) Phase Measurements at Slough

The variations in the phase of the ionospheric wave during the months of February, March, June and July, 1953, and September, 1952, were found to be similar to the earlier results of Hopkins and Reynolds,[5] and so they need not be reported in detail. If the phase changes are interpreted as changes in the height of reflection, the fall in height began one to two hours before sunrise and was substantially complete about three hours after sunrise. The rise from day to night levels extended over the period from three

to four hours before sunset to two to three hours after sunset. Changes in height were somewhat greater than those quoted by Hopkins and Reynolds; the range was about 17 to 19 km, the higher values occurring in winter, in contrast with the earlier results.

(4.2) Derivation of Conversion Coefficients from Bearings

Results of previous measurements of the reflecting properties of the ionosphere have been presented in terms of the reflection and conversion coefficients, which are, respectively, the amplitude ratios of the vertically and horizontally polarized components of the reflected wave to the incident wave—assumed vertically polarized. The information to be derived from the direction-finding observations is related mainly to the conversion coefficient, since the errors are caused by the horizontally polarized component of field, but by using both bearing and amplitude data, reflection coefficients can also be derived (see Section 4.6).

The conversion coefficients are derived in two stages—first the calculation of the horizontal field at the aerial and then the interpretation of this field in terms of the reflecting properties of the ionosphere.

If the voltage induced in the direction-finding aerial system by the horizontally polarized component of field is ρ times that induced by the vertically polarized component, and if the voltages differ by a phase angle ϕ, the error θ and ellipticity r are given by

$$\tan 2\theta = \frac{2\rho \cos \phi}{1 - \rho^2}$$

$$r = \rho \sin \phi \, (1 - 2\rho^2 \cos^2 \phi)^{\frac{1}{2}} \quad \text{when} \quad \rho^2 \ll 1$$

It follows that, when ρ is less than $0\cdot5$, ρ^2 equals $r^2 + \tan^2 \theta$ to a close approximation. For convenience the term 'error' will be used to denote values of $\tan \theta$ rather than θ.

From the value of ρ, which relates to the fields at the aerial, the conversion coefficient of the ionosphere is deduced on the assumption that the ground is a perfect reflector, that the effective height of reflection is 80 km and that the transmitter radiation follows a cosine law in the vertical plane. These assumptions were the same as those used in previous work,[5,9] and they lead to the expression $\frac{1}{2}\rho \operatorname{cosec}^2 i \sec i$ for the conversion coefficient, where i is the angle of incidence.

One difficulty in obtaining the true values of ρ lies in the distortion of the field due to imperfections of the site. Some allowance can be made for the 'fixed' site errors by examination of the symmetry of the error curves during their cyclic excursions, but it is difficult to measure accurately the small day-time errors caused by polarization effects. However, the time variations of error are relatively unaffected by the fixed errors, and the horizontally polarized field strength can be obtained from the amplitude of the cyclic error variations at sunrise and sunset. This method is analogous to the polar-plot method of studying amplitude and phase variations, described by Bracewell.[9] It is noteworthy also that, whereas instantaneous values of error and ellipticity are influenced by the vertically polarized component of the sky wave, when this is appreciable compared with the ground wave, the amplitudes of the oscillations are relatively unaffected.

(4.3) Conversion Coefficients for the Rugby–Slough Path

Conversion coefficients derived from the Slough measurements for four periods in 1953 and 1954 are plotted in Fig. 5. As the main interest was in night-time errors, the time scale has been centred on midnight. The small dots were obtained from smoothed curves of error and ellipticity, making allowance, where possible, for fixed errors. The large dots were derived

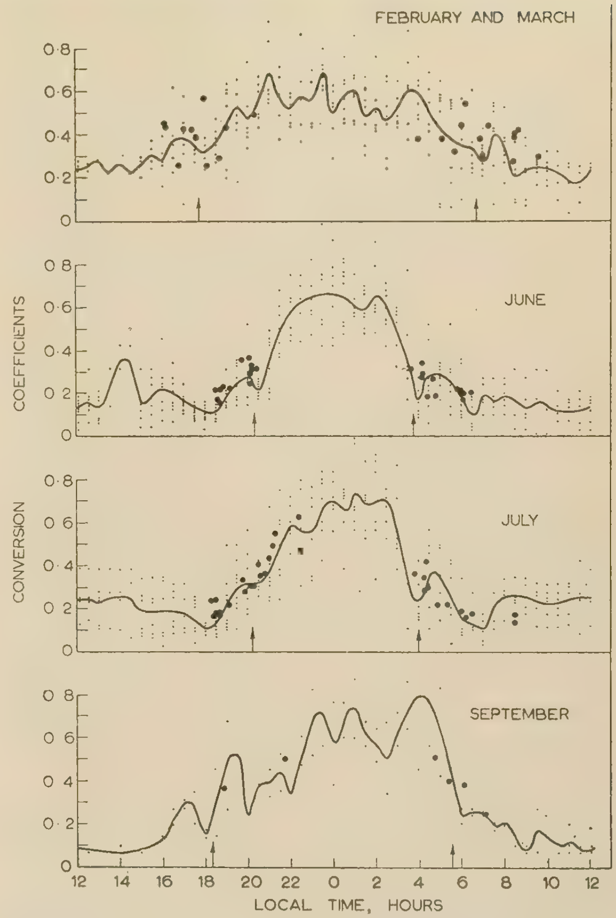


Fig. 5.—Conversion coefficients for the Rugby–Slough path.

• From total errors.
● From oscillations of error curves.

Curves show mean values derived from total error data in intervals of half an hour. Sunset and sunrise times are indicated by arrows.

from the oscillatory parts of the error curves and should be relatively free from the effects of fixed errors. Smooth curves have been drawn through average values for successive time intervals.

The general form of the curves is in agreement with those of Hopkins and Reynolds.[5] The summer curves show well-defined changes at sunrise and sunset, while the winter curves show more gradual changes. During the periods of fully established night conditions, average values lie mainly in the range $0\cdot55$–$0\cdot75$. These are significantly higher than those quoted for Cambridge[4] ($0\cdot40$–$0\cdot55$) as were those of Hopkins and Reynolds. In comparing the results of the different workers it may be noted that, although information is recorded in different forms, the assumptions used to derive conversion coefficients are the same in all cases and affect the results to the same extent.

The possible dependence of propagation conditions on direction, at very low frequencies, has been discussed for many years. The view was expressed by Naismith[10] in 1931, as a result of field-strength measurements, that the propagation of signals from GBR was dependent on direction. Later experiments at Cambridge[11] did not support this view, but the more recent work of Hopkins and Reynolds suggested that there may be a directional effect, although other possible reasons for the observed differences required consideration. The experiments to be described in the following Sections were designed to establish

more reliably whether there was a real difference between the two propagation paths and, if so, to investigate possible reasons.

(4.4) Simultaneous Measurements at Slough and Cambridge

To ensure that the discrepancies between results at different sites were not caused in any way by the use of different types of instrument, a direction-finder was operated near the Cambridge ionospheric recording equipment for two days. For comparison the records of the Cambridge equipment were expressed in the form of the apparent bearing of the transmitter. It was found that the variations in apparent bearing were almost identical on the two equipments, but that there was a fixed difference between them of about 5°. Similar but smaller differences have been noticed on other occasions; they were attributed to the presence of cables and other conductors on the site, possibly with a small contribution from errors in setting up the equipment.

In view of these 'fixed' errors, comparisons between different equipments and sites were subsequently based entirely on the variations in apparent bearings rather than on their absolute magnitudes. Since the variations of which use was made occurred mainly during the transition period between daylight and night conditions, no particular significance is attached to the actual values of conversion coefficients; it is the ratios of the values obtained simultaneously at different sites which are of interest.

The comparison is facilitated by the fact that the path length between the ground and once-reflected ionospheric wave at Cambridge differs by half a wavelength from the corresponding value at Slough. The error curves therefore tend to be similar, but with a reversal in sign.

Some of the Cambridge data to be discussed were derived from direction-finders and others from the equipment operated by the Cavendish Laboratory. No distinction will be made between the two sources.

The results obtained in simultaneous measurements at Slough and Cambridge are expressed in Table 2 in terms of the ratios of the derived conversion coefficients. An 'observation' is a measurement of conversion coefficient from a single excursion of the error (or ellipticity) curve from a positive to a negative peak or vice versa, for example from 0625 to 0730 hours in Fig. 1.

Table 2

RELATIVE CONVERSION COEFFICIENTS AT SLOUGH AND CAMBRIDGE

Date	No. of observations	Coefficient at Cambridge divided by that at Slough	
		Range	Average
March, 1954 ..	2	0·79–0·98	0·89
July, 1954	6	0·66–0·90	0·76
September, 1954 ..	9	0·50–0·90	0·75
August, 1955 ..	9	0·53–0·84	0·71

Overall average ratio, 0·75.

The coefficients for Cambridge are consistently below those for Slough, and the average is considerably lower. This could be caused by a real difference in the reflecting properties of the ionosphere or could be an apparent effect caused by neglect of waves arriving after more than one ionospheric reflection, or by incorrect assumptions regarding the modes of propagation of the ground and ionospheric waves or the radiation pattern of the transmitter. Further experiments were carried out to investigate these possibilities.

(4.5) Comparisons between Reception at Slough, Cambridge and Penn

Slough and Cambridge are, respectively, 108 and 90 km from Rugby, and this difference may have two effects on the results. In the first place it is necessary to assume a law of attenuation of the ground wave (an inverse distance law has been used). Secondly, the conversion coefficients are derived on the assumption that only the ground wave and the horizontally polarized component of the once-reflected ionospheric wave are significant. The method used should largely eliminate the effect of the vertical component of this wave, but the waves which are reflected more than once at the ionosphere will influence the results in a manner dependent on the distance from the transmitter. To investigate possible effects from these multiple-reflection waves, measurements were made at Penn, which is the same distance as Cambridge from Rugby but in the direction of Slough. Two sites were, in fact, used—at 86 and 90 km from Rugby.

The ratios of conversion coefficients are shown in Table 3.

Table 3

RELATIVE CONVERSION COEFFICIENTS AT SLOUGH AND PENN

Date	No. of observations	Coefficient at Penn divided by that at Slough	
		Range	Average
March, 1954 ..	4	0·73–1·04	0·91
July, 1954	3	0·83–1·17	1·05
August, 1955 ..	18	0·81–1·37	1·02

Overall average ratio, 1·01.

There is much closer agreement with Slough at Penn than at Cambridge, and most of the data were obtained simultaneously at all three stations. (The much larger number of observations for Penn in August, 1955, resulted from the use of both the error and ellipticity curves.) It was therefore concluded that the differences between Slough and Cambridge were not caused by the neglect of multiple reflections.

From the amplitude measurements at the various stations an attempt was made to assess the field strength of the ground wave and to check the inverse-distance relationship. These measurements were rather inaccurate owing to the presence of the ionospheric waves, to variations in the output of the transmitter and to difficulties of preserving an accurate calibration of a mobile direction-finder operated from unstable power supplies. Nevertheless the average field strengths at the various stations conformed to an inverse-distance law to within a few per cent, with the exception of one station where data were obtained for only a 24-hour period. Also in July, 1954, the field strengths measured at Cambridge were lower than those at Penn, which would tend to give higher conversion coefficients at Cambridge. The coefficients were, in fact, considerably lower, as can be seen from Tables 2 and 3.

The possibility was also considered that the assumptions made regarding the radiation pattern of the transmitter were incorrect. However, Hopkins and Reynolds[5] conclude from a study of the transmitting aerial that the ratio of the horizontally to vertically polarized components of the transmitted wave, at the appropriate vertical angles, is only of the order of 10% at the most. It also seems unlikely that the radiation pattern of the transmitter would differ significantly from the assumed cosine law in the vertical plane.

The differences between Slough and Cambridge therefore appeared to be related to the reflecting properties of the iono-

sphere, and it was decided to investigate the polarization of the waves.

(4.6) Reflection Coefficients and Polarization

By considering the variations in amplitude in conjunction with the bearing data it is possible to estimate the reflection coefficient of the ionosphere, i.e. the amplitude ratio of the vertically polarized component of the reflected wave to the incident wave. The derived values are less accurate than those of the conversion coefficient for several reasons:

(*a*) Field-strength comparisons are inherently more difficult than those involving direction-finding observations, owing to the need for preserving an accurate calibration of the equipment. The difficulties are particularly noticeable with mobile equipment.

(*b*) Information about reflection coefficients is derived from small variations in the total field strength; the main contribution, at the distances under consideration, being from the ground wave.

(*c*) There was some evidence that the output from the transmitter was not constant. Similar changes of field strength at more than one station often appeared to be caused by changes of transmitter power.

In addition, the multiple-hop waves affected the field-strength measurements to at least the same extent as they affected the bearings.

The method of measuring the relative values of the reflection and conversion coefficients was to compare the variations in the bearing error and ellipticity with the corresponding variations in amplitude. In the records of August, 1955, from which the most complete data were obtained, it was observed that the maxima and minima of amplitude coincided in time, during the sunrise and sunset periods, with the minima and maxima of the ellipticity curve (see, for example, 0600 to 0700 hours G.M.T. in Fig. 1). The horizontally and vertically polarized components of the ionospheric wave were therefore approximately in phase quadrature, and it could be shown that the direction of rotation of the field vector was anti-clockwise, looking in the direction of propagation, in agreement with previous workers.[4]

One result of this phase relationship is that the horizontally polarized component has little effect on the maximum and minimum amplitudes, which are determined by the vertically polarized, once reflected wave, with a small contribution from the multiple-hop waves. It is therefore valid to assume that, as the ellipticity changes from its maximum positive to maximum negative value, the corresponding amplitude change, from minimum to maximum gives a direct indication of the pick-up from the vertically polarized component of the ionospheric wave. The magnitudes of the changes are summarized in Table 4.

Table 4

COMPARISON OF VERTICALLY AND HORIZONTALLY POLARIZED COMPONENTS

	Slough	Penn	Average between Slough and Penn
Range of $\Delta V/\Delta H$ ..	0·37–1·18	0·23–1·09	0·39–1·08
Average values ..	0·64	0·66	0·64
Corresponding V/H	0·54	0·58	0·55

$$\frac{\Delta V}{\Delta H} = \frac{\text{Fractional change in amplitude}}{\text{Corresponding change in ellipticity}}$$

V, H = Vertically and horizontally polarized components of the ionospheric wave.

There is considerable scatter in the ratios, some of which was thought to be caused by changes in transmitter power. As the changes caused by variations in the height of the reflecting layer are in phase opposition at the two sites, while those due to transmitter power variation are in phase, some reduction in scatter can be achieved by averaging the changes at Penn and Slough, as shown in the last column.

The ratios of V/H reported for Cambridge from early experiments[11] were about 1·5, but later papers[4,12] quote a value of unity at all times. Using the latter value the coefficients for Slough and Cambridge may be summarized by the following process:

Slough

From Fig. 5 the average value of the conversion coefficient around midnight in summer is 0·65.

From Table 4 the polarization ratio, V/H, is 0·55.

The reflection coefficient is therefore 0·36.

Cambridge

From Table 2 the conversion coefficient is three-quarters of the value for Slough or 0·49.

From previous work at Cambridge the polarization ratio is unity.

The reflection coefficient is therefore 0·49.

Considering the power in both the vertically and horizontally polarized waves, about half the incident power is reflected along both the Slough and Cambridge paths, but there is a difference in polarization which is sufficient to explain the differences in the conversion coefficients. This conclusion has some support from the results of Bracewell[9] for a distance of 200 km. Along two paths, which were similar in direction to the Slough and Cambridge paths discussed here, the polarization ratios, V/H, were 0·6 and 1·0, respectively.

(5) CONCLUSIONS

The primary purpose of the experiments was to measure directly the polarization error on c.w. transmissions and to deduce the errors which would occur on atmospherics under similar conditions. It is concluded that there is no major difference in the influence of the unwanted components of field on the bearings, but under conditions where the c.w. signal is subject to large bearing errors, atmospherics, owing to their transient nature, often produce traces too complex for a bearing to be read.

Apart from this consideration the errors are substantially the same for both types of signal. Errors are largest in the range 200–400 km, where they have median values of the order of 10° by day. At night they are much larger, the median rising to a maximum value of about 30° at 300 km, but decreasing rather rapidly with further increase in distance. At 1 000 km, median errors of about 2° may be expected by day and by night.

In the operational observations on atmospherics the practical importance of polarization errors will be reduced by several factors. First, as mentioned above, no bearing may be read if the trace is complex. Secondly, if the errors are large no fix may be obtained from the bearings. Thirdly, the largest errors occur at about 300 km and the positional error in the fix may not be large compared with the extent of the storm. Nevertheless it might be possible to achieve a worth-while improvement in accuracy by weighting the bearings according to their probable errors.

The polarization-error measurements provided a means of studying the reflecting properties of the ionosphere and comparing the findings with those from previous work. It appears that the differences between the earlier results from the paths Rugby–Cambridge and Rugby–Slough are real, and are due to a difference in the polarization at the two receiving sites. This result suggests that the polarization of the wave reflected by the ionosphere is dependent on the direction of propagation, but it is possible that the observed effects are due partly to the radiation of a horizontally polarized field by the transmitter. This field is

likely to be small compared with the vertically polarized field, but it has not yet been shown to be negligible.

(6) ACKNOWLEDGMENTS

The author wishes to acknowledge the co-operation of Mr· A. L. Maidens of the Meteorological Office and Dr. K. Weekes of the Cavendish Laboratory, who provided facilities for experiments and data from their own observations. Messrs. V. A. W. Harrison and C. Clarke of the Radio Research Station carried out much of the experimental work.

The work described above was carried out as part of the programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(7) REFERENCES

(1) Ockenden, C. V.: 'Sferics', *Meteorological Magazine*, 1954, **83**, p. 137.

(2) Adcock, F., and Clarke, C.: 'The Location of Thunderstorms by Radio Direction-Finding', *Journal I.E.E.*, 1947, **94**, Part III, p. 118.

(3) Horner, F.: 'The Accuracy of the Location of Sources of Atmospherics by Radio Direction Finding', *Proceedings I.E.E.*, Paper No. 1709 R, November, 1954 (**101**, Part III, p. 383).

(4) Bracewell, R. N., Budden, K. G., Ratcliffe, J. A., Straker, T. W., and Weekes, K.: 'The Ionospheric Propagation of Low- and Very-Low-Frequency Waves over Distances less than 1 000 km', *ibid.*, Paper No. 1123 R, May, 1951 (**98**, Part III, p. 221).

(5) Hopkins, H. G., and Reynolds, L. G.: 'An Experimental Investigation of Short-Distance Ionospheric Propagation at Low and Very Low Frequencies', *ibid.*, Paper No. 1600 R, January, 1954 (**101**, Part III, p. 21).

(6) Clarke, C., and Harrison, V. A. W.: 'Low Frequency Direction Finder, Twin-Channel Cathode-Ray Design', *Wireless Engineer*, 1954, **31**, p. 109.

(7) Lutkin, F. E.: 'Directional Recording of Radio Atmospherics', *Journal I.E.E.*, 1938, **82**, p. 289.

(8) Lugeon, J.: 'Le radiogoniograph de la Station Centrale Suisse de Météorologie et son utilisation pour la prévision du temps', *Annales de la Station Centrale Suisse de Météorologie*, 1939.

(9) Bracewell, R. N.: 'The Ionospheric Propagation of Radio Waves of Frequency 16 kc/s over Distances of about 200 km', *Proceedings I.E.E.*, Monograph No. 34 R, April, 1952 (**99**, Part IV, p. 217).

(10) Naismith, R.: 'Short Distance Observations on Long-Wave Phenomena', *Journal I.E.E.*, 1931, **69**, p. 875.

(11) Budden, K. G., Ratcliffe, J. A., and Wilkes, M. V.: 'Further Investigations of Very Long Waves reflected from the Ionosphere', *Proceedings of the Royal Society*, A, 1939, **171**, p. 188.

(12) Straker, T. W.: 'The Ionospheric Propagation of Radio Waves of Frequency 16 kc/s over Short Distances', *Proceedings I.E.E.*, Monograph No. 114 R, January, 1955 (**102 C**, p. 122).

Abstract No. 2323
Feb. 1957

SCOTTISH CENTRE: CHAIRMAN'S ADDRESS

By Professor F. M. BRUCE, M.Sc., Ph.D., A.Inst.P., Member.

'POST-GRADUATE RESEARCH: ITS OBJECTS AND SOME ACHIEVEMENTS'

(ABSTRACT *of Address delivered before the* SOUTH-EAST SCOTLAND SUB-CENTRE *at* EDINBURGH *2nd October, the* SOUTH-WEST SCOTLAND SUB-CENTRE *at* GLASGOW *3rd October, and the* NORTH SCOTLAND SUB-CENTRE *at* DUNDEE *11th October and at* ABERDEEN *12th October*, 1956.)

In these days of large industrial research and development organizations, what might be defined as research in the classical manner is now given a prefix such as 'fundamental', 'long-term' or 'academic', and the experience of post-graduate work leading to a higher degree is invaluable to the small proportion of graduates who exhibit the particular aptitudes required for such work. They will later be concerned with seeking to establish reliable theories or data for imperfectly understood physical phenomena, all the emphasis being placed upon the search for true facts, and without any defined objective such as always lurks in the background of 'industrial' or 'applied' research. They must develop a good sense of logic in planning their work, a power of analytical judgment, and that absorbing interest which alone can produce the initiative required.

Some seven years ago, the Governors of the Royal College of Science and Technology approved a programme for the establishment of a research school in the Department of Electrical Engineering, and this has now been operating effectively for some years. In particular, very comprehensive facilities for research at high voltages have been made available.

For this Address, I have chosen to deal, in a general way, with two of the main programmes of research—the mechanism of the spark discharge in air, and power follow-current phenomena.

Spark Studies

The introduction of high impulse voltages, and suitable oscillographs for recording the very rapid change of voltage with time, provided a new means for studying the mechanism of sparkover. It was soon found that gaps exhibited time-lag phenomena, as instanced by breakdown occurring on the tail of the wave and not at the crest. For a given gap and impulse waveform, there is a range in the crest values of the impulses applied to the gap over which the frequency of breakdown for a number of applications varies from 0 to 100%. The zero level could be taken as the threshold of breakdown for the given conditions, and an increased frequency of breakdown required a higher voltage (or an 'over-voltage') for which the point of breakdown moved from the tail towards the crest of the wave, indicating a reduction in the time-lag of the gap.

More fundamental studies of time-lag phenomena were undertaken with direct voltages, giving rise to the division of time-lag into statistical and formative periods. The former can be eliminated by adequate irradiation, leaving the latter as characteristic of the breakdown mechanism. It was found that the formative time could be reduced, apparently without limit, if sufficient over-voltage was applied.

Whether with direct or impulse voltages, the time-lags were of the order of microseconds or very much less, and these low values could not be explained by the processes involved in the Townsend theory as then appreciated, unless at low values of $p \times d$. For these reasons, the *Kanal* or streamer theories were proposed as the breakdown mechanism[1] for the high-pressure discharge, and it was believed that transition from the Townsend to the streamer mechanism occurred at some critical value of $p \times d$.

For reasons previously published,[2] I was not satisfied with this transition in the mechanism or the qualitative arguments of the alternative theory, and further investigation of it was one of the first items on our programme. The uniform field, which is the basis for studies of this type, is produced between parallel plane electrodes having the edges suitably curved to ensure that breakdown, within the working range of spacings, always takes place in the uniform-field region. At excessive spacings, of course, the maximum surface gradient is developed at the curved edges, giving non-uniform field discharges. There have been various designs of electrode contour to produce these conditions, and it is perhaps natural that an atmosphere of some rivalry has appeared in the past, and may well do so in the future.

In fact, the shape of the edges is immaterial, provided that it can be demonstrated that the sparks have the characteristics of breakdown in a uniform field. The edge contour does have secondary effects such as determining clearance required from surrounding objects, overall dimensions, and maximum permissible spacing. For my part, I continue to use the Stephenson profile, with which I have worked for more than 20 years.[3] This is the only type of electrode for which it has been established that, for alternating voltages, the sparkover-voltage/spacing characteristic can be expressed by a single equation to an accuracy within $\pm 0 \cdot 1\%$. The equation applies to electrodes of any overall dimensions and is independent of polarity. I am glad to announce that our work in Glasgow has already established that the equation applies equally to direct and impulse voltages of either polarity. These characteristics are to be expected in truly uniform-field breakdown.

The time-lag for these uniform-field gaps was investigated with stable direct voltages. An approach voltage some 2% below the sparking voltage is applied, and on this is superimposed a rectangular pulse voltage—by measuring this voltage separately, the pulse magnitude can be expressed very accurately as a percentage of the total. The magnitude of the pulse is increased by small increments until breakdowns occur, the time from application of the pulse to breakdown giving a measure of the time-lag. This work has now been extended to gap spacings up to 4·0cm, the results being consistent with data previously published[4] for smaller spacings. The time-lag/over-voltage curves are repeatable, and independent of electrode dimensions provided that the field is uniform. At low over-voltages (about $0 \cdot 1\%$), time-lags of the order of hundreds of microseconds are obtained, and a plot of time-lag/spacing for constant values of over-voltage yields a series of straight lines passing through the origin. These are characteristics of the Townsend theory as now understood,[5] and have been determined for values of $p \times d$ up to 3000, or 15 times as high as once thought, and the range is being extended.

The direct-voltage time-lag data reveal the limitations inherent in using impulse voltages only, for such studies. In that case, the 0% criterion has to be taken as the threshold of sparking, but because the peak voltage only lasts for times of the order of microseconds, when at or near to the crest value of the waveform, this impulse datum is really a value of over-voltage sufficient to ensure that the time-lag has already been reduced to a value lying within the duration of the peak voltage of the waveform. The true threshold voltage is the direct-voltage datum, and will have a lower value. Using a 0·2/240 impulse wave (this being

Professor Bruce occupies the Chair of Electrical Engineering at the Royal College of Science and Technology, Glasgow.

an even closer approximation to direct voltage than the standard 1/50 wave), data have already been published[4] to account quantitatively for the apparent discrepancy between impulse and direct-voltage time-lag values.

Photographic studies of spark discharges have been used in the College for some years, the first series being based on the method of suppressed discharges described by Torok. The technique was modified by using a fixed loop to control the reflection time, but inserting a series gap, the time-lag of which was varied to control the time to suppression in the main gap. This was applied to sparking between spheres, and revealed a number of important characteristics in the development of the discharge.[6]

A uniform-field gap, especially when highly 'over-volted', reveals auto-suppression. This is due to the very small formative time-lag under these conditions, so that several channels are likely to develop at the same time, and the first to be completed then suppresses the others. By using a camera technique giving two orthogonal views, the spark position relative to the electrodes and to each other could be identified. One of the most interesting photographs was that showing a rectangular step in the path of the spark.[6]

Another means for studying the mechanism of sparkover is in the measurement of the currents due to ion transport immediately prior to breakdown. Using a technique involving a long time-constant, it appeared that there was a steady build-up of current in an irradiated gap, and a sudden build-up, at breakdown, in an unirradiated gap. A technique using a short-time-constant circuit revealed the fine structure of the current build-up as a series of pulses. A sensitive photomultiplier can pick up the emission of light from some of these pulse discharges, and we have recently been able to take simultaneous records of the current and light pulses.

Work of this type is now being planned for both uniform and non-uniform field breakdown, and at higher voltages. Uniform-field electrodes for working up to 1 000 kV are at present being made.

The influence of atmospheric humidity on the breakdown voltage of an air-gap is a matter of immediate importance to high-voltage engineers, but the data available are very limited. It seemed that the high consistency of the uniform-field gap made it a suitable means for detecting even such small changes in breakdown voltage as may be caused by variations in humidity, and a programme of research to this end has been in progress for some years. For gaps of the order of 1–2 cm we have observed an increase in sparkover voltage with humidity for alternating, direct and impulse voltages. It remains to determine the possible contribution made to this effect by phenomena occurring at the electrode surface and in the air, and it is most desirable that the work should be carried to higher voltages, and include cases of divergent-field breakdown at voltages of practical interest. The present work is therefore to be regarded as introductory to work on a much larger scale.

The present situation is that we now have experimental techniques that exceed, by an order of magnitude, the sensitivity given by changes in breakdown voltage as an indication of a change in the breakdown mechanism of a spark-gap. Indeed, we have found it impossible to use some of these because of the slight variations that occur in the ambient atmosphere of the laboratories. It is also necessary to check the validity of theory over a wider range of atmospheric conditions than that afforded even by the Glasgow weather. For these reasons we are at present installing a controlled-atmosphere chamber some 9 ft in height and 5 ft in diameter in which it is hoped to work at up to 200–300 kV in a stabilized atmosphere. Ancillary apparatus gives a pressure range of 0–2 atm absolute, temperature control from -5°C to $+30^\circ$C, and the full range of possible humidity values. Provision has been made for the application of dual photographic and other means of observation. In addition to providing more stable conditions for basic studies, it will be possible to produce conditions of icing, or low temperature and pressure, on such apparatus as insulators or on aircraft equipment.

Power Follow-Current Investigations

When flashover occurs on energized high-voltage equipment, a power arc will, in most cases but not all, be established in the path of the initial spark, and cause a short-circuit. The power follow-current carried by the arc has then to be interrupted. Work has been proceeding in the College for some years on the study of power arcs initiated by the impulse breakdown of a gap. A synthetic power source comprising a tuned *LC* circuit was first used with rod-gaps of up to 10 cm and impulse voltages up to 200 kV, as a result of which the various types of phenomena were classified and the important parameters identified. This synthetic power source can readily be modified, by altering the *LC* values, to give a range of frequencies and power source impedances. The impulse breakdown of the gap occurs under conditions simulating the application of a surge voltage at the instant when the power-system voltage is at a maximum, whether of like or opposite polarity to that of the surge. It was found that the critical condition for the development of a power arc was determined by arc-glow transitions which occurred at currents of the order of 1 amp. Once the arc condition has been established, the short-circuit current that develops will be determined by the characteristics of the power source, but the critical condition is well within the range of laboratory equipment.

The investigation was then extended to include the effect of the point-on-wave at which the impulse was applied, using either high-voltage transformers as conventional power sources or the synthetic technique modified so that the voltage on the storage capacitor was varying at the instant of application of the impulse. Gaps ranging from 10 cm to 15 in were studied, and similar results were obtained with the two power sources. The most favourable condition for a power arc to develop occurs when the impulse is applied while the power-source voltage is still rising to its peak value.

Similar techniques could be developed for work on transformer insulation, and the synthetic circuit offers a means of producing a controlled amount of power follow-energy in the path of an impulse breakdown when testing transformers, so that the exact point of failure can subsequently be located.

Acknowledgments

I cannot name all the individual members of staff and research students who have contributed to the work described, but their names will be identified with the specialized papers that have been, or will be, published. I would, however, acknowledge the assistance received from the electrical industry in various ways, and from the Carnegie and Sir James Caird Trusts, and the Central Electricity Authority, in the award of research scholarships.

References

(1) JONES, F. L.: *British Journal of Applied Physics*, 1954, **5**, p. 49.
(2) BRUCE, F. M.: *Journal of the British Institution of Radio Engineers*, 1951, **11**, p. 121.
(3) BRUCE, F. M.: *Journal I.E.E.*, 1947, **94**, Part II, p. 138.
(4) AKED, A., BRUCE, F. M., and TEDFORD, D. J.: *British Journal of Applied Physics*, 1955, **6**, p. 233.
(5) DUTTON, J., HAYDON, S. C., JONES, F. L., and DAVIDSON, P. M.: *ibid.*, 1953, **4**, p. 170.
(6) BRUCE, F. M.: *Endeavour*, 1954, **13**, p. 61.

Abstract No. 2331
Feb. 1957

SOUTH MIDLAND CENTRE: CHAIRMAN'S ADDRESS

By C. J. O. GARRARD, M.Sc., Member.

'EDUCATION'

(ABSTRACT *of Address delivered at* BIRMINGHAM, 1*st October*, 1956.)

By its Royal Charter, The Institution is required 'to facilitate the exchange of information and ideas on Electrical Science and Engineering'. It is therefore one of our duties to concern ourselves with the education of electrical engineers and thus with technical education in general. There are four main fields in which this work is done, namely through our requirements for membership and the Institution Examination; through the Graduate and Student Sections; through the opportunities for further education provided by our ordinary meetings and discussions; and by advice given to Government, local authorities and other bodies concerned with education.

The Examination Regulations

The recent revision of the Examination Regulations has raised not a little the required standard, particularly in mathematics and physics, and has widened the syllabus in a number of subjects. It is to be hoped that, in applying the new standards in the schools and colleges, regard will be paid to the supreme importance of the elucidation of principles and the cultivation of the ability to learn from continued experience, as well as to the teaching of facts. Much difficulty in technical and scientific education might be avoided if more attention were given to what appear to be deficiencies in the elementary teaching of English and arithmetic—subjects which must continue to be the foundation of all more advanced studies.

The Institution, too, would do well to concern itself with the excessive specialization, not only in technological but also in humane studies, both in secondary schools and in universities. In electrical engineering courses, the growing unbalance between heavy- and light-current work is causing concern. The remedy in both cases appears to lie more in a general recasting of the course work than in mere additions to an already overloaded syllabus.

Graduate and Student Sections

The Graduate and Student Sections provide opportunities, not only for self-help by their 16 000 members, but for the senior members of The Institution to play their part in the education of the profession. It is essential that every opportunity be taken to cultivate that relation of pupil to teacher that is in danger of being lost, and of encouraging young engineers to acquire experience of responsibility at the earliest possible age; this in many cases may be more beneficial than continued academic training or post-graduate research. As a corollary to this we should encourage the provision of academic training for men who have already spent some time in industry.

If a man stays on to do research after taking his first degree, it is most desirable that the subject for investigation should be such that it can be grasped and carried through in a reasonable time by a single man or a group of two or three, who themselves can do the necessary thinking, planning and manipulation.

Meetings

The most striking development in the day-to-day work of The Institution is the continued increase in the number and diversity of its meetings, an increase to which the length of the year and the growing pressure of members' own work must presumably sooner or later set a limit.

The creation of Specialized Sections and Groups, while it has encouraged the multiplication of papers, symposia and conventions, has, by decentralization, provided to some extent at least the means for coping with the consequent greatly increased work of organization and has supplied the necessary audiences. The need to maintain adequate liaison between the resulting more or less different interests, however, lays upon the officers of The Institution a burden which, if it should be still more increased, may give rise to concern.

In organizing our meetings we should perhaps distinguish between two related but distinct functions of The Institution. One of these is to provide an exchange of information on the frontiers of electrical science. This it seems might best be done through *ad hoc* group meetings for the discussion by specialists of papers of more limited interest. At these meetings the papers could be taken as read, and the whole of the available time devoted to discussion, unless there appeared to be some advantage in an oral explanation of the subject.

The other function of The Institution is to give its members a general picture of the state of electrical engineering and to be a forum for the discussion of matters of general interest. Meetings for this purpose could often with advantage be based on lectures rather than papers; the lecture is often a more convenient means of summarizing and presenting a subject in a general way than is a formal paper.

At all our meetings, we should strive to increase the breadth and freedom of discussion by limiting the number of prearranged contributions, more strictly enforcing any necessary time limit on individual speakers and observing the existing rule against the reading of prepared statements.

Work outside The Institution

To-day almost all public discussion of technical and scientific education is concerned directly or indirectly with the shortage of scientists and engineers and with means to remove the shortage. The Institution is deeply concerned in these discussions, both in the persons of individual members and corporately.

At the personal level there are three things that might be said. The first is that there are grounds for thinking that the scarcity of technical staff is being aggravated to some extent by wrong and wasteful use of the people that are available. There is a great need for the application of work-study techniques to brain-work. Much time and time and trouble have been expended in applying time-and-motion study, methods development and similar techniques to the manual work of relatively plentiful and relatively unskilled workpeople. Some of this effort could well be diverted to studying the work of draughtsmen, designers and other technical workers.

Second, the fundamental difficulty seems to be a shortage of science teachers. If we want more we must see that the status of a teacher is at least as good as that of a doctor or a lawyer.

The third point is that if, as citizens or engineers or employers, we demand more and better scientific education we must not grumble when, as taxpayers, we are given the bill.

It is most essential that local authorities should be persuaded

Mr. Garrard is with The General Electric Co., Ltd.

to have the same liberality of outlook in their dealings with the colleges under their care as the University Grants Committee have in their dealings with the universities. The authorities should provide the money, see that it is not wasted, and beyond that interfere as little as possible in the affairs of the colleges.

Corporate action of The Institution will no doubt continue to be taken in supporting and improving the courses that are being set up for the recently instituted Diploma in Technology. One must agree that there is much force in the view that effective training can be carried out only in strongly staffed colleges offering a wide range of subjects; that a mediocre standard attracts only mediocre students and results in awards that are of little value. There is substance, however, in the view that the need is great and that help should be taken where it can be got.

The greatest need for many years will be for competent staff; here members of The Institution can help, either by part-time teaching themselves, or by making it as easy as possible for those of their staff who are willing to take up part-time teaching.

On a more formal basis, collaboration between industry and the colleges will be of the greatest value. An example is the sandwich courses now in being or being organized in many places. These are intended to supplement the existing university courses and to provide for those boys who do not get places in the universities a better engineering education than can be given by an apprenticeship with perhaps one day and an evening or two per week in college. It is felt, incidentally, that few boys will be able to cope with the new Institution Examination with only one day a week at college.

A sandwich course consists of periods of about six months spent alternately in the works and in the college for a total time of 5 years. The courses are designed to give the students a sufficiently thorough grounding in the basic disciplines of mathematics, physics, chemistry and so on, to enable them to think independently about their problems rather than merely to follow routine and rule of thumb. On the other hand, it is intended to give them, by the time they get into industry, some specialized competence and so enable them to establish themselves fairly rapidly as competent and authoritative engineers in the eyes, not only of their contemporaries, but also of men, sometimes their seniors, who have not had the same educational advantages.

While the active participation of industry in educational effort is wholly desirable, two possible tendencies should be kept in mind. One is that of diverting effort away from education, which should be directed towards the good of the individual and of the community in general and is the concern of private individuals and of the State, and towards training, which is properly directed towards the needs of a firm, a service or an industry, and should in the main be their responsibility.

The second tendency is that of accentuating the already grave shortage of good teachers and good administrators in the public sector of education by transferring many of the more capable people into industry. It might be supposed that the shortage of scientifically trained people would automatically set up conditions leading to an increase of supply. This unhappily does not seem to be true. In fact the shortage of technical men in industry is draining science teachers from the schools and thus creating an even greater potential shortage.

In education, as in many other things, one of our greatest needs is clearly to define and where necessary to delimit our objectives and then to concentrate upon their attainment a due portion of our scanty resources.

Even inside The Institution there may be some need for concentration of effort. We all welcome the steadily increasing bulk and intensity of its work. This is a sign of vitality, energy, public spirit and enthusiasm. One sometimes wonders, however, whether we are not approaching a time when we shall have to exercise some restraint in undertaking new commitments, especially in fields of a less essential character, in order that those that exist may be adequately discharged. While the work of The Institution is important and essential, our collective daily duties in the profession and the industry are even more important. In fact they are an irreplaceable foundation of the commerce and life of the nation.

We hear a great deal from time to time about education for leisure; there is doubtless need to teach people to spend their spare time more profitably, and leisure is necessary if the arts and graces of life are to develop at all. Nevertheless there is a danger that this propaganda may deflect us from what I think to be the truth, namely that we in this country are much less in danger from excessive leisure than from being obliged to work excessively hard.

If one looks at economic and technical developments at home and abroad, it is difficult to avoid the conclusion that in the not very distant future we are likely to have to work a good deal harder than we have in the past. More and more we are being turned back upon our own resources of materials and manpower. More and more our manufactures have to compete abroad with locally made goods and with the exports of countries which have a lower standard of living than ours; that is, whose inhabitants work harder and longer for less material reward than we do.

Added to all this is the fact that advances in medicine and hygiene are causing an explosive increase in the world's population, an increase at the moment of about one per second. This increase as it continues will exercise a pressure on food and other resources to which our children and grandchildren in these islands will be ever more exposed. The development of atomic energy, of automatic control and automatic production may give us some help. No one who has examined the question quantitatively, however, appears to think that the total effect will be much more than marginal for very many years.

We have to reckon with the possibility that the era of relative ease and prosperity through which we have been passing may prove to have been, not a prelude to even better things, but a passing phase. We should therefore be on safer ground, I think, if we tried to gear our education and related activities, such as those of this Institution, to preparation for harder work rather than for more leisure.

In so doing, however, we should not forget that the real need of the world is for fresh ideas. The fundamental notions upon which the whole of our scientific civilization is built are now quite old, some of them several thousand years old. We are involved in working out the ever-increasing mass of detailed consequence that has flowed from them. We are in fact in some danger of diverting effort from fundamental inquiry to short-range projects and of assuming that the only important aim of education is the production of material benefits.

No civilization can long survive unless it so orders its affairs that some at least of its citizens can pursue learning for its own sake, with open minds stimulated only by curiosity. Experience has shown during all the ages that it is such men and they alone who achieve the epoch-making discoveries that ensure future progress.

534.846

Paper No. 2297 R
Mar. 1957

ACOUSTICS OF LARGE ORCHESTRAL STUDIOS AND CONCERT HALLS

By T. SOMERVILLE, B.Sc., F.Inst.P., Member, and C. L. S. GILFORD, M.Sc., F.Inst.P., Associate Member.

(*The paper was first received* 2nd *July, and in revised form* 24*th October*, 1956.)

SUMMARY

In order to establish design criteria for orchestral studios, which should be similar in performance to good concert halls, an extensive investigation has been in progress for many years.

The effects of shape on the subjective acoustic qualities of a large enclosure are here examined with reference to a large number of concert halls and music studios, and a comparison is made, in particular, between concert halls of the traditional type and those which have been built during the last few decades. The former were generally of rectangular plan with walls and ceilings overlaid with ornamentation, whereas most recent designs have fan-shaped plans, reflecting canopies and comparatively smooth surfaces.

Measurements of sound levels in different parts of concert halls during orchestral concerts show that, for a given sound level in the neighbourhood of the platform, the intensity at the back of the hall is no greater in halls with fan-shaped plans and reflectors than with the traditional rectangular shape. In the former case, the gain in the intensity of the first few reflections which results from the shape of the hall is offset by a reduced reverberant sound level.

The authors conclude that the modern fashion of directing the early reflections towards the back of a concert hall, although it may improve the hearing of speech, has an adverse effect on the quality of music.

(1) INTRODUCTION

A considerable portion of B.B.C. programme time is devoted to orchestral music, and the Corporation therefore maintains a number of studios to accommodate orchestras of between 60 and 100 performers. In the interests of realism it is necessary that programmes produced in these studios should have the characteristics associated with orchestras in concert halls, and for this reason the B.B.C. has studied the acoustics of many concert halls to establish acoustical standards for its own use. It is the purpose of the paper to describe the conclusions reached as a result of investigations made in many large studios and concert halls.

In recent years several new concert studios and concert halls have been built, both in this country and throughout Europe. It has now been possible to make detailed acoustic surveys of all the modern studios and concert halls in Great Britain and also to visit corresponding places of interest on the Continent, though without making instrumental measurements. The various points of interest will be discussed in the paper in the light of present-day experience. Most modern designs have been based on the belief that it is necessary to adopt fan shapes, splays and reflectors to obtain satisfactory acoustics. Comments by one of the authors[1] in 1949 when comparing the Liverpool Philharmonic Hall and St. Andrew's Grand Hall, Glasgow, may therefore be of interest. The conclusions at that time were as follows:

> These experiments have been of great interest, because there is a widespread belief that good acoustic conditions can be obtained by reflecting sound forward into the auditorium. There are several studios or concert halls in which this has been done, but the Philharmonic Hall is the only example in this country. The performance of St. Andrew's Hall is of particular interest, because there are no large reflecting areas near the orchestra and yet the hall fulfils all the requirements for good acoustics. It seems, therefore, that the projection of sound is not necessary, but it still remains to be proved whether good acoustics can be obtained in fan-shaped halls if proper precautions are taken to produce adequate dispersion.

The article was based on detailed measurements in two halls only and a survey of a few others. At the time there was widespread interest in the design of concert halls, with special reference to the Royal Festival Hall, Free Trade Hall and Queen's Hall sites. Fan-shaped plans were much favoured by those who were involved in these projects. For the Royal Festival Hall, a rectangular plan was adopted, but many of the effects of fan shapes were obtained by means of splays. Since that date, three large concert halls have been built in this country alone, and several B.B.C. music studios varying in size from 30 000 to 220 000 ft^3 have been rebuilt or acoustically treated.

The paper gives results of investigations made mainly in Great Britain. Objective measurements are confined to information about shape, construction and dimensions and to the acoustic parameters which are universally recognized, such as reverberation time and sound-level distribution. Details of an acoustic criterion, based on objective measurements, which gives good agreement with subjective assessments, are also given. The more general subjective assessments are based on opinions collected over a number of years by the authors from publications, newspaper criticisms and conversations with musicians and listeners. There will inevitably be disagreement with these views, which are unlikely to be accepted universally, but they are believed to be a fair representation of informed opinion.

The more detailed descriptions of acoustic characteristics are the judgments of the authors and their colleagues, and recordings demonstrating many of these effects are in existence.

Before proceeding to describe the recent investigations it may be desirable to define the terms commonly used to describe the subjective qualities of a concert hall or studio.

(2) GLOSSARY OF ACOUSTIC TERMS

Balance.—The loudness relationship between different groups of musical instruments in an orchestra, as heard from a point in the auditorium or by means of a microphone.

Bass masking.—The masking of instruments such as strings and woodwinds by the low-frequency instruments such as brass and percussion. It is observed in enclosures which give undue weight to the low-frequency instruments.

Coloration.—A characteristic timbre imparted by the acoustics of the enclosure. It is often caused by undamped normal modes, or resonances in the structure, which cause reradiation at specific frequencies after the exciting sound has ceased.

Deadness.—The opposite of 'liveness'. A characteristic of enclosures in which the reverberation time is very short, or of points in an enclosure where the ratio of direct to reverberant sound is high.

Definition.—That quality which enables all the parts in an orchestral work to be heard clearly.

Diffusion (*formerly dispersion*).—Describes uniformity in distribution of sound energy in an enclosure.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Somerville and Mr. Gilford are with the British Broadcasting Corporation.

Echoes.—Discrete reflections from surfaces so situated that the reflection arrives after the wanted sound has died away.

Flutter echoes.—A rapid multiple echo of even rate. It is usually caused by parallel reflecting surfaces between which sound is reflected in a periodic manner.

Liveness.—The term is applied (*a*) to enclosures in which the reverberation time is high, or (*b*) to a point in an enclosure where the ratio of direct to reverberant sound is low.

Pitch changes.—Rise or fall of pitch due to frequency changes in the reverberant sound as it dies away in an enclosure. (Not to be confused with the purely subjective slight fall in pitch which is always heard as the intensity of sound of fixed frequency is reduced.)

Scattering.—The effect of irregularities in the surfaces; by distributing sound, it tends to improve diffusion.

Singing tone.—A property which enables an enclosure to respond easily to any frequency and appears to be related to the manner in which the sound dies away.

Slap back.—Echoes from the rear surfaces of an enclosure.

Standing-wave system.—An interference pattern characterized by stationary nodes and anti-nodes.

(3) CHARACTERISTICS OF OLD HALLS

Most of the concert halls built during the latter half of the nineteenth century resembled to some extent the Gewandhaus at Leipzig. Hence these halls are often referred to as the 'Leipzig' type. They have side and rear balconies, and the orchestra is situated on a flat platform behind which there is raked seating for the choir. Fig. 1 shows St. Andrew's Hall, Glasgow, which

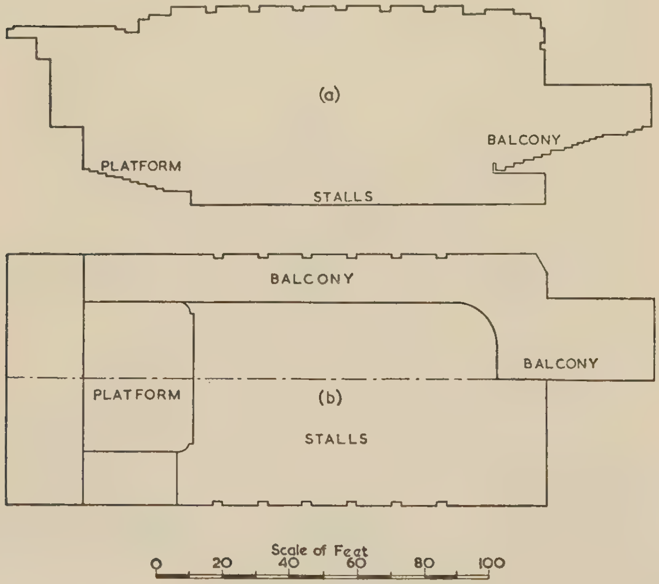


Fig. 1.—St. Andrew's Hall, Glasgow.
(*a*) Long section.
(*b*) Plan.

is a typical example. As these halls were designed for other purposes in addition to music, the floor is usually flat. In the early days much wood was used in the internal decoration, but towards the end of last century it became fashionable to employ lath and plaster with 'stick and rag' construction for the ornamentation, which was usually very complex. Such irregularities scatter the sound waves and produce good diffusion, although at that time this was not recognized to be an essential requirement for good acoustics. Because of the seating placed behind the orchestra, there is considerable absorption, thus making the region where the brass and percussion instruments are normally situated more dead than the front of the platform where the strings perform. The methods of construction of floors, ceilings and balconies resulted in great variation in the frequencies of the structural resonances, so that the sound absorption from this source was evenly distributed and colorations due to resonances were avoided. Good British examples of Leipzig-type halls were the Queen's Hall, London, the old Free Trade Hall, Manchester, the old Colston Hall, Bristol, the old Philharmonic Hall, Liverpool, and St. Andrew's Grand Hall, Glasgow. Unfortunately the only remaining example of these old halls is St. Andrew's Hall, all the others having been destroyed by fire. Therefore it has received detailed study in an endeavour to find the reason for their outstanding acoustic properties.

(3.1) St. Andrew's Hall, Glasgow

Most of the older halls had reasonably long reverberation times. Fig. 2 shows the reverberation times* of concert halls

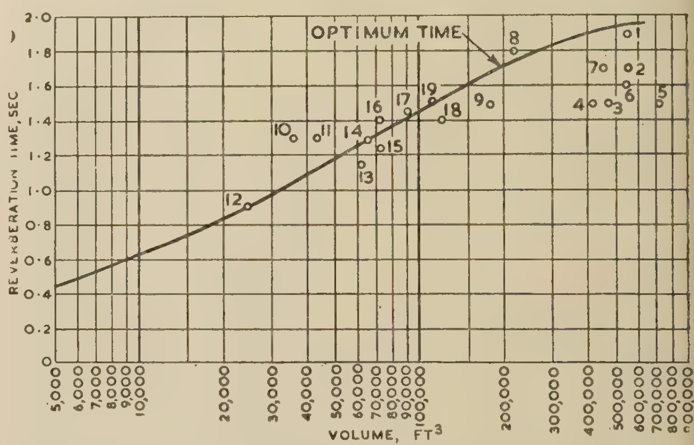


Fig. 2.—Relation between optimum reverberation time and volume.

Reverberation times with orchestra, and audience where applicable.

1. Glasgow, St. Andrew's Hall.
2. Edinburgh, Usher Hall.
3. Liverpool, Philharmonic Hall.
4. Copenhagen, Concert Hall.
5. London, Royal Festival Hall.
6. Manchester, Free Trade Hall.
7. Bristol, Colston Hall.
8. London, Maida Vale, Studio 1.
9. Glasgow, Studio 1.
10. Swansea, Studio 1.
11. Cardiff, Charles St. Hall.
12. Glasgow, Studio 2.
13. London, Maida Vale, Studio 2.
14. London, Maida Vale, Studio 3.
15. Belfast, Studio 1.
16. Birmingham, Carpenter Road Studio.
17. Manchester, Milton Hall.
18. London, Broadcasting House, Concert Hall.
19. London, Farringdon Hall.

and concert studios discussed in the paper, plotted as a function of volume. In St. Andrew's Hall this is 1·9 sec with the audience and orchestra present. In addition, the variation between different parts of the auditorium is small, thus indicating excellent diffusion. The roof construction is entirely in timber supporting a deeply coffered ceiling of plaster, the outline of which may be seen in the long section in Fig. 1. The walls are lined with lath-and-plaster panelling in a varied design, except near floor level in stalls and balconies, where wood panelling is used for protective purposes. The floor is oak strip on joists over a crypt used for storage. This construction results in

* The figures given are the mean values for the frequency range 125 c/s–4 kc/s

distribution of the structural resonances throughout the low-frequency region, with the result that the hall is notably free from colorations. Bass notes are very clear and unmarred by pitch changes during the decay of sound. Another characteristic of this hall is the fact that, although the reverberation time is long, it is possible to hear all the individual parts, even in loud passages. This characteristic is well described by the music critic of *The Times* as follows:[2]

> St. Andrew's Hall is notoriously generous in its acoustics but it does not blur outlines and the fact that the texture was also clear as well as warm was a certificate of the orchestra's purity of intonation.

It may therefore be concluded that, if other essential features have been incorporated in the design, a reasonably long reverberation time is not a disadvantage. These features are:

(*a*) Good diffusion.
(*b*) A reverberation-time/frequency characteristic that does not rise at low frequencies.
(*c*) Local absorption behind the orchestral platform to keep the powerful instruments under control.

All these requirements have been fulfilled in St. Andrew's Hall, which has a richness of tone quite absent from all modern halls. One disadvantage of the long reverberation time is that the hall, although good, is not ideal for the hearing of speech. Parkin[3] has stated that the reverberation time is obtained at the expense of echoes from the rear and that to cure them absorption would be necessary. So far as the authors are aware the normal users of this hall have never complained of echoes, a judgment with which they agree. (This is confirmed by the fact that in all post-war subjective comparisons, e.g. that of Parkin, Scholes and Derbyshire,[4] this hall has been ranked either first or second of British halls.) In any case, the area of absorbing material which would be required to suppress echoes would be too small to reduce the reverberation time appreciably. It would indeed be possible to employ diffusing surfaces instead, which would distribute the energy over a wide angle without decreasing the reverberation time.

While trying to find reasons for the excellent acoustics of St. Andrew's Hall, a survey of the literature was carried out, and the discovery was made that the acoustic consultant for this hall was the architect who had designed the old Liverpool Philharmonic Hall completed in 1849. This architect, John Cunningham, had an extensive practice in Liverpool until he retired in 1872. Because the Liverpool Philharmonic Hall was considered to have excellent acoustic properties, Cunningham was engaged to act in a consulting capacity for St. Andrew's Hall. This was opened in 1877, but Cunningham died before its completion. It is unfortunate that he left no description of the details of the acoustic design of these halls, for, although it is often said that the acoustics of the old concert halls were largely a matter of chance, it may be significant that Cunningham was responsible for two halls in succession both of which were good.

The success of these two concert halls might be explained if the architect copied known good designs, although in this connection Caird Hall, Dundee, which is similar in design to St. Andrew's Hall, is not successful, probably because its length is too great for the other dimensions.

Several good concert halls built during the latter half of the nineteenth century have been mentioned, but this does not imply that all halls built during this period were good: the converse is, in fact, true. This was the period of rapid industrial expansion and consequently of extensive building activities including the construction of many halls throughout Great Britain. Most of them were not built specifically for music and were, in fact, not satisfactory for this purpose.

(3.2) Usher Hall, Edinburgh

In the early part of this century the design of concert halls began to differ from the Leipzig type, one well-known example being the Usher Hall, Edinburgh, opened in 1914. All the orchestral concerts at the Edinburgh International Festival are given here. This hall has a horseshoe plan with raked seating in the stalls and two balconies (see Fig. 3). Behind the orchestra

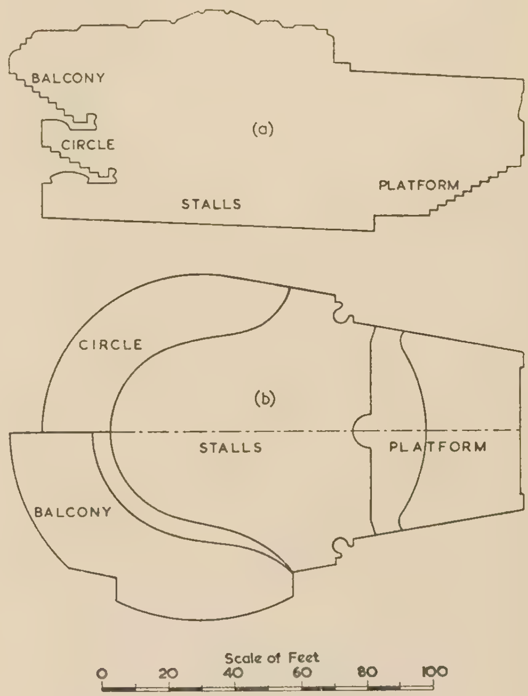


Fig. 3.—Usher Hall, Edinburgh.
(*a*) Long section.
(*b*) Plan.

are choir seats, but the side walls are close to the orchestra since the balconies do not go round beside the orchestra as in the Leipzig type. The hall is smaller in volume than St. Andrew's Hall, and as the seating accommodation is greater it has a reverberation time of 1·7 sec, which is less than is desirable. Hence the tonal quality is, by comparison, a little harsh although the hall is quite good. There are some bass colorations apparently due to structural resonances in the stage. The extensive hard surfaces near the sides of the orchestra sometimes cause accentuation of instruments which happen to be placed in the vicinity. There is less scattering than in good Leipzig-type halls and consequently less diffusion, and this would explain some observed variation in acoustics between different positions in the auditorium.

(4) CHARACTERISTICS OF MODERN HALLS

It has not been possible to find out the origin of the present-day belief that modern concert halls should resemble cinemas and that it is necessary to adopt a fan-shaped plan and to surround the orchestra by reflectors to project the sound energy on to the audience. With small rooms rectangular plans may cause flutter echoes, but this phenomenon has not been observed by the authors to be of importance in any large concert halls. The reason usually given for a fan-shaped plan is that it improves the sound level at the back of the hall. The fashion appeared to start with the Salle Pleyel in Paris, which is an early example of a hall in which sound projection was attempted.

(4.1) Salle Pleyel, Paris

The features of the Salle Pleyel have been described by Bagenal and Wood,[5] the basic principle being that sound should be reflected as directly as possible on to the audience, which in this hall is seated in raked stalls and two balconies. Fig. 4 shows the long section and plan of the hall. According to Andrade[6] the result, so far as speakers are concerned, is eminently satisfactory because the reverberation time is short and echoes are absent. However, the hall has a poor reputation for musical acoustics—to such an extent that it is seldom used for orchestral concerts. The reasons for this will be discussed in connection with modern halls in Great Britain which the authors have been able to investigate in detail. Although the behaviour of the Salle Pleyel is not outstanding, the design principles are supported authoritively in most modern textbooks. Furthermore, those principles have been applied in varying degrees in most modern designs.

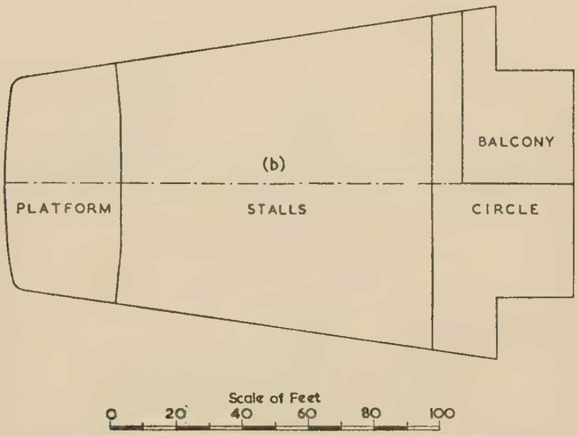
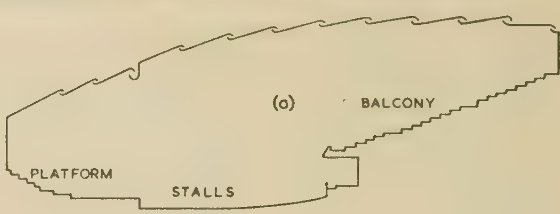


Fig. 4.—Salle Pleyel, Paris.
(*a*) Long section.
(*b*) Plan.

(4.2) Philharmonic Hall, Liverpool

The earliest example of a modern hall in Great Britain is the new Liverpool Philharmonic Hall, completed in 1939 to replace the old concert hall which had been burnt down. This hall is fan-shaped and the walls and ceiling are designed to act as reflectors so that the sound energy is directed from the orchestra on to the audience. A long section and plan are shown in Fig. 5. In any concert hall the audience is the main source of absorption, and it is therefore not surprising that, because much of the sound reaches the audience after very few reflections and consequently cannot contribute to reverberation, the hall is dead for its volume. The volume is 476 000 ft^3 and the reverberation time is about 1·5 sec with audience. A detailed study of this hall has been made in comparison with St. Andrew's Hall, Glasgow,[1] so that only a summary is necessary here. The tonal quality is a little harsh, owing to the short reverberation time, and the distribution of sound is somewhat patchy because of poor diffusion. It is found that the powerful instruments tend to mask the others but that this effect can be reduced if the choir seats behind the orchestra are occupied. A good feature is that there are no seats under balconies, which always present difficulty because of screening. The performance of the hall, while not comparable with the old Liverpool Philharmonic Hall, is satisfactory.



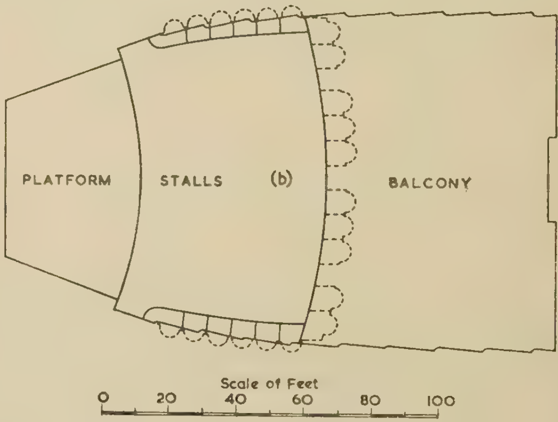


Fig. 5.—Liverpool Philharmonic Hall.
(*a*) Long section.
(*b*) Plan.

(4.3) Concert Studio, Copenhagen

An interesting example of modern design is a large concert studio built in Copenhagen in 1942 by the Danes, during the German occupation. The long section of this studio (see Fig. 6) is very reminiscent of the Salle Pleyel, but in plan it is more fan-shaped. The volume is 420 000 ft^3, and the reverberation time is 1·5 sec, which is short. In this studio the rear walls were made curved, and it was therefore necessary to apply a thick layer of absorbing material to prevent echoes, although this obviously increases the absorption which is already too high. Because of the deadness and the lack of diffusion, the tonal quality is harsh and the acoustic characteristics vary considerably from place to place. It is impossible to hear all the instruments in the orchestra in heavy passages, the more powerful instruments near the reflectors being those which predominate. For broadcasting a multi-microphone technique is therefore essential in order to enable all the parts to be heard.

(4.4) Concert Hall, Gothenburg

A modern concert hall having an acoustic performance very similar to that of the Liverpool Philharmonic Hall has been

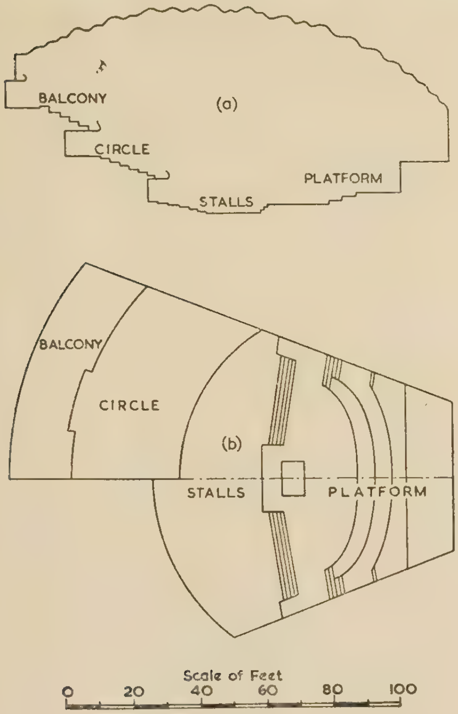


Fig. 6.—Concert Studio, Broadcasting House, Copenhagen.

(*a*) Long section.
(*b*) Plan.

built in Gothenburg. It also is designed to reflect sound from the orchestra, but, whereas in Liverpool the interior surfaces are constructed in plaster, wood panelling is used in Gothenburg. There is no balcony but the rear seating is steeply raked. The treatment of the back wall is interesting because the architect avoided the use of absorption to prevent echoes by sloping the wall so that reflections from it would come down on the seats towards the rear. Unfortunately, in certain areas marked echoes can be heard from this surface. In this hall also, hearing is difficult and several microphones are necessary when broadcasting.

(4.5) Royal Festival Hall, London

The first concert hall to be built in Britain after the war was the Royal Festival Hall, adequate descriptions of which are to be found in the literature.[7, 8] The long section and plan are shown in Fig. 7. Its designers decided that a rectangular plan should be employed in preference to a fan shape, but although the structure is rectangular, splays, reflectors over the orchestra, and a shaped ceiling have been used to project the sound on to the audience, so that the effects of a fan shape are obtained. This means that the hall is dead for the volume. The reverberation time with an audience is 1·5 sec, and the volume is 760 000 ft^3. Its designers were aware of the advantages of diffusion in concert halls, and therefore, although the general shape of the ceiling was designed to project sound forwards, it was corrugated to give diffusion. Unfortunately, this is only effective in longitudinal directions and there is little diffusion transversely. A feature of the design which caused much comment was the provision of boxes on the side walls. According to the designers these were intended to provide diffusion, in addition to accommodating a larger audience. From the observed effect of boxes in opera houses on the Continent, the authors believe that little diffusion is obtained from this source but that there is a considerable increase in absorption produced by such cavities containing audience and draperies. If the side walls had been made

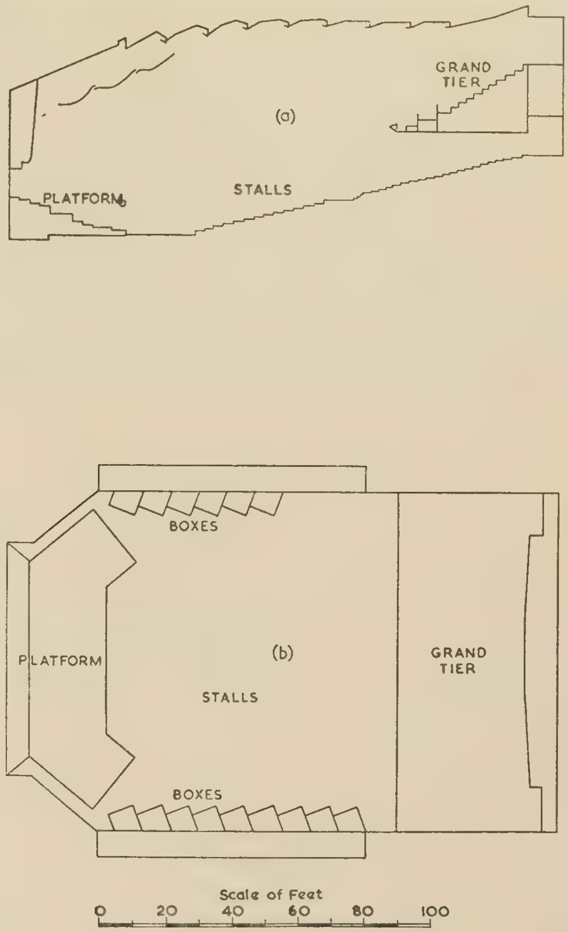


Fig. 7.—Royal Festival Hall, London.

(*a*) Long section.
(*b*) Plan.

diffusing, but without the increase in absorption provided by the boxes, it is probable that this hall would have been more reverberant than it now is.

The Royal Festival Hall is very good for the hearing of speech because of its deadness and the absence of prominent echoes. It is also good for chamber music and for the Mozart-size orchestra. As the size of the orchestra increases a true ensemble becomes more difficult to obtain, and for a full-size symphony orchestra, conditions, in the authors' opinion, are not entirely satisfactory. Press comments on the hall in the first eighteen months after its completion were divergent.[8] The majority considered the 'fullness' or 'resonance' to be satisfactory, but a minority amounting to about 40% considered the hall too dead.

There was similarly a fairly considerable minority of unfavourable reports on blend and balance; one of these,[9] with which the authors are in general agreement, is of interest:

> Two kinds of music seem to suffer the most; the elusive atmospheric piece, and the sumptuous late Romantic scores of Wagner, Strauss and Rachmaninoff. . . .
>
> It is doubtless one and the same fault which prevents these categories of music from having their proper effect. Probably the reverberation period of the hall is still just too short; possibly the steep slope of the orchestra and auditorium encourages the brass and percussion to drown the strings in the valley between. Whatever the cause it is quite astonishing how differently music of a fully saturated texture can sound in a suitable auditorium such as the Prinzregenttheatre in Munich or the Concertegebouw in Amsterdam.
>
> By the way, I must protest at the assumption that those of us who are not quite satisfied . . . are displaying a taste vitiated by

long experience of the Albert Hall. . . . I can think of no famous auditorium with the peculiar dryness of tone which is to be felt at every big climax in the Festival Hall. There seems to be a growing concensus of opinion about this fault. . . .

(4.6) Free Trade Hall, Manchester

The new Free Trade Hall, Manchester, was the second large hall to be built in Britain after the war. Although intended to accommodate the Hallé Orchestra, the necessity to use it also for other purposes placed restrictions on the design as a concert hall. The original hall of the Leipzig type, which was similar in performance to the old Liverpool Philharmonic Hall and St. Andrew's Hall, Glasgow, was destroyed by bombing. It was necessary, for economic reasons, to increase the seating in the new hall, and for this purpose, in addition to side balconies, a circle and a balcony were added, as shown in Fig. 8. To

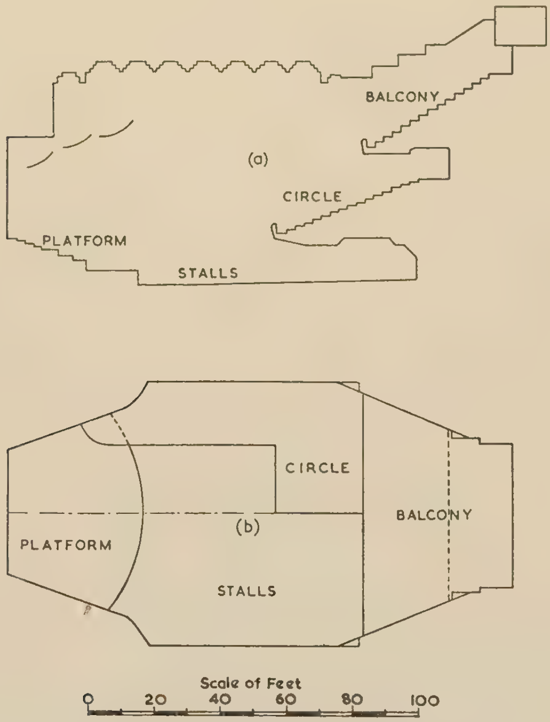


Fig. 8.—Free Trade Hall, Manchester.
(*a*) Long section.
(*b*) Plan.

counteract screening by the balconies it was decided to fit splays behind the orchestra and reflectors overhead, but the ceiling was made flat and coffered because a reflecting ceiling was not considered to be necessary. The projection of sound is therefore not so efficient as in the other modern halls described, but even so the reverberation time is only 1·6 sec with audience in a volume of 535 000 ft^3. Consequently, although the reverberation time is shorter than is desirable, the tonal quality is less harsh and the acoustic characteristics are more satisfactory than those of the Royal Festival Hall. There is, however, still difficulty caused by the reflectors over-emphasizing the powerful instruments, and hearing conditions vary somewhat throughout the auditorium. Because of screening, hearing is only fair at the back of stalls, circle and balcony.

(4.7) Colston Hall, Bristol

The Colston Hall was rebuilt in 1951, after destruction by fire, using the original outer shell. The auditorium plan, shown in

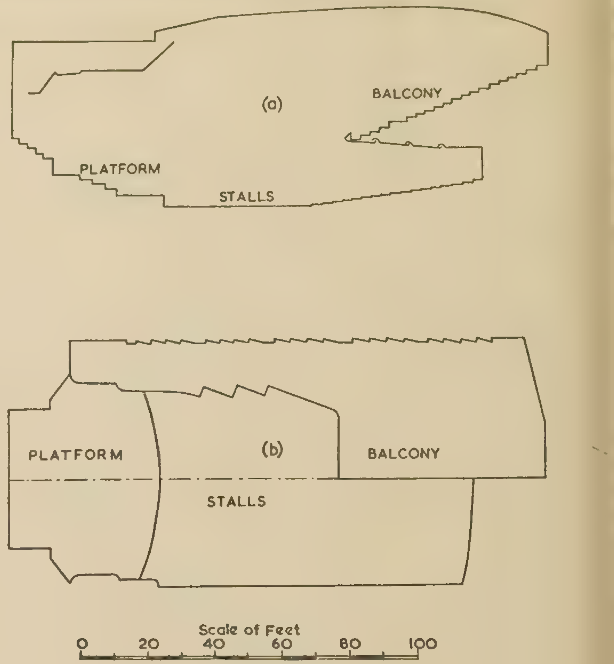


Fig. 9.—Colston Hall, Bristol.
(*a*) Long section.
(*b*) Plan.

Fig. 9, is rectangular apart from slight modifications of the back wall, and the ceiling is slightly concave in both directions. No attempt was made to break up the flat ceiling areas either with lighting rolls as in the Royal Festival Hall, or with coffering as in the Free Trade Hall. Lighting is by hanging fittings, and the only breaks in the ceiling surface are the ventilation outlets. In this hall the canopy differs from those in other post-war halls. The part of it immediately over the central stage is horizontal, with the object of reflecting sound back to the orchestra itself in order to enable the performers to hear the balance between their own instruments and the orchestra as a whole. The remainder of the canopy reflects sound into the audience, but the total sound energy thus directed is much less than in other post-war halls. The height of the auditorium ceiling is greater in relation to the volume than is the case with most concert halls. To give a convenient representation of this relationship, we may calculate the ratio of the overall height to the geometric mean of length and breadth. Comparative figures are as follows:

Liverpool Philharmonic Hall	0·34
Royal Festival Hall	0·39
St. Andrew's Hall	0·49
Colston Hall	0·53
Free Trade Hall	0·60

The orchestral platform has high risers to ensure good direct sound to the main seating area. The upper parts of the side walls are panelled with veneered chipboard arranged in a sawtooth plan, providing bass absorption and scattering, and the walls below the balcony are panelled in solid wood with a deeply modelled profile.[10] High-frequency absorbers are sparingly used only where thought necessary to prevent long-path echoes.

The reverberation time of this hall is 1·7 sec between 500 and 1 000 c/s with full audience, and just over 2 sec when empty. The former figure is higher than that for the other two post-war halls in spite of the fact that its volume (450 000 ft^3) is considerably less. The tonal quality is accordingly fuller than in either of those halls, and this has been achieved without the production of

echoes. The definition is good but not, perhaps, exceptional; with the hall empty there is a tendency for the tympani and brass to be prominent, but this is less noticeable with a full audience.

(4.8) General Characteristics

In all these modern halls, except perhaps the Colston Hall, the reverberation time is shorter than the normally accepted values for their volumes. The result is therefore harsh tonal quality which always goes with deadness, but is further accentuated by the fact that the sound decay is uneven as a result of insufficient diffusion—an inevitable consequence of the projection of the sound in one direction. All of them, to a greater or less extent, are affected by bass masking, caused by the fact that the strings are away from reinforcing surfaces, whereas the noisy brass and percussion instruments are close to the reflectors and mask the weaker instruments in loud passages. Bass masking can also occur in halls where the reverberation time increases at low frequencies, but as this fault is recognized by British designers, most modern halls have been given bass absorption sufficient to keep the characteristic down at low frequencies. Because the projection of energy towards the rear is efficient in many of these halls, the rear surfaces have been covered with absorbent material to prevent the effect known as 'slap back'. Only in the concert hall in Gothenburg has the rear wall been inclined to prevent echoes without the need for extra absorption.

(5) TYPICAL DESIGN CRITERIA

As a result of the investigation of many large studios and concert halls in this country and abroad, and while the authorities concerned were still considering the design of the Royal Festival, Free Trade, and Colston Halls, design criteria were formulated as a basis for the acoustic design of B.B.C. orchestral studios.

B.B.C. experience is that a concert hall which is good for broadcasting with a single microphone will always be good for direct listening, but the converse is not necessarily true. This is brought about by the fact that the binaural hearing of the normal member of a concert audience enables many of the faults which would be deleterious to broadcasting to be ignored, and it therefore follows that many concert halls which are accepted by an audience are unsatisfactory for broadcasting. In some cases, on the other hand, by suitable arrangement of microphones, it has been possible to obtain good broadcast results from halls which have bad reputations with concert audiences.

The criteria used by the B.B.C. in concert-studio design may therefore be of some interest, since concert studios are intended to reproduce the acoustic characteristics of good concert halls.

(5.1) Shape

Until 1939 all B.B.C. music studios were built in the traditional rectangular form. The interruption in the studio building programme between 1939 and the end of the war prevented the building of any large orchestral studios. When it was possible to consider rebuilding, surveys of British and Continental concert halls led to the opinion that fan shapes, which cause the sound to be reflected away from the orchestra, produce unsatisfactory conditions. As reflectors and concave ceilings also have the same effect they have been avoided in all modern designs by the B.B.C.

Since the end of the war the B.B.C., because of continued restrictions on building, has from time to time acquired small halls for use as concert studios. Some of these with ecclesiastical origins have pitched roofs, so that the ratio of height to length and breadth is greater than that normally adopted in concert studios. In most cases these halls have good acoustics and give an acoustic impression of size which is greater than the dimensions would indicate. The concert studio in Glasgow is also unusually high, and here again the subjective impression of spaciousness is very noticeable. Generally speaking, there is no evidence that the dimensions of large studios or concert halls require to have any special relative proportions. It is sufficient that the plan should not be too square and that the height should be more than about one-half the smaller plan dimension.

(5.2) Absorption

In large enclosures such as concert studios and halls the only absorption necessary is at low and medium frequencies because air attenuation and the audience produce more high-frequency absorption than is usually desired. In most recent B.B.C. designs the low- and medium-frequency absorption has been obtained by the use of membrane absorbers, as described elsewhere.[11] These are shown in Fig. 10 in the large B.B.C. studio at Maida Vale. The rectangular form of these absorbers is also an advantage in obtaining good diffusion. In accordance with the normal practice in the traditional concert hall, absorption in the form of choir seats is placed behind the orchestra—as shown in the Figure. Fig. 11 is a photograph of the concert studio in Glasgow. As there is no seating for a choir in this studio it has been necessary to apply absorption to the vertical surfaces immediately behind the orchestra to produce a similar effect. Originally this studio did not have absorption behind the orchestra, and its introduction has therefore provided a useful confirmation of the fact that absorption behind the brass and percussion instruments is highly desirable. To enable the members of the orchestra to hear themselves playing, parallel surfaces on either side have been left reflecting.

(5.3) Diffusion

The need for excellent diffusion in concert studios has already been discussed. There is evidence to indicate that the most important surface in a concert studio on which to apply diffusers is the ceiling, but this does not mean that the wall surfaces should not also receive treatment. This opinion is based on the fact that in many of the old concert halls of the Leipzig type the ceiling was more elaborately ornamented than the walls, and in the concert studio in Glasgow this also applies, apparently with good results. Although considerable diffusion is necessary, it is an advantage to place small parallel reflecting surfaces on either side of the orchestra, as was often done in old concert halls, because this enables the performers to hear each other easily.

The methods of obtaining diffusion in Maida Vale are interesting. It would have been desirable to put most of the scattering surfaces on the ceiling, but this old studio, built in 1935, was of such a construction that the heavy mass of the absorbers could not be carried by the ceiling and therefore it had to be accommodated on the walls. A few light diffusers, however, were placed on the ceiling to break up some large flat areas.

(6) SOME TYPICAL ORCHESTRAL STUDIOS

In this Section a brief description will be given of a few of the larger B.B.C. music studios which have been built or acoustically retreated since the war.

(6.1) Maida Vale, Studio 1

Some of the design details of Maida Vale, Studio 1, have been described in Section 5. This studio, being built inside an existing shell, was restricted in height to 25 ft. Its volume is 220 000 ft^3 and its mean reverberation time with an orchestra is 1·8 sec. The main roof structure tends to absorb fairly strongly between

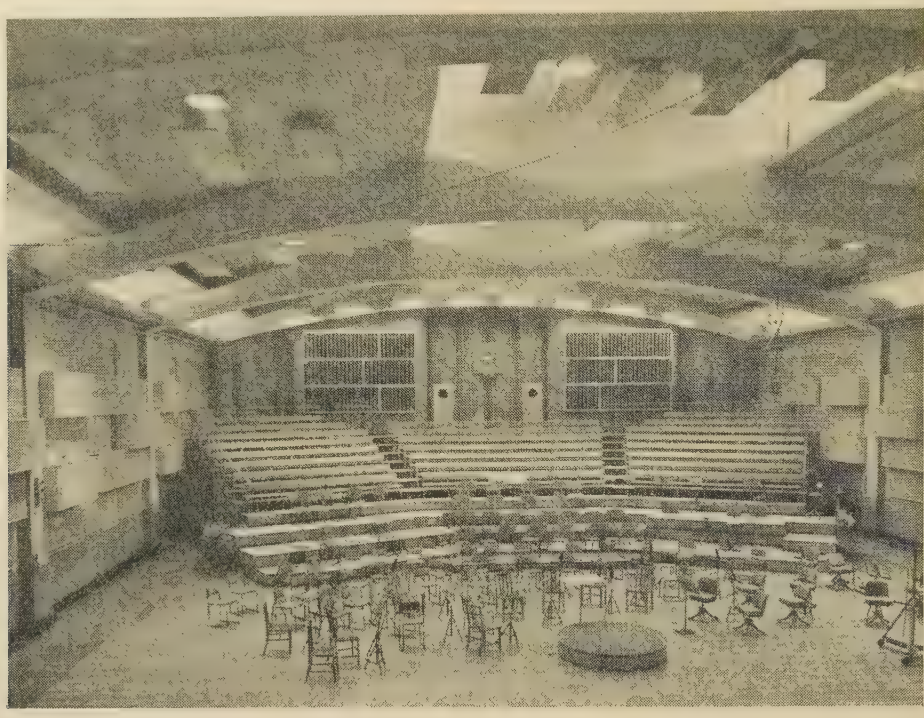


Fig. 10.—B.B.C. Concert Studio, Maida Vale.

100 and 150 c/s, resulting in a slight subjective deficiency of bass. Diffusion is effective and definition is good. Tonal quality is usually regarded as excellent.

(6.2) Glasgow, Studio 1

Like the Maida Vale studio, Glasgow, Studio 1, is rectangular but though the volume (180 000 ft^3) is less, it has the greater height of 40 ft. The reverberation time is 1·7 sec, sloping gradually up from a broad minimum of 1·4 sec at 125 c/s. The distribution of absorbing material has been described in Section 5; there is little scattering except on the ceiling. It will accommodate an orchestra of 65–70 players without any adverse effect on the acoustics. The tonal quality is rather brighter than that of Maida Vale, owing to a better maintenance of the reverberation time at high frequencies. The definition is extremely good, and the studio gives an aural impression of considerable size which may be connected with the unusual height in relation to the length of 80 ft and width of 57 ft. The main construction is of stone and heavy brickwork, but low-frequency absorption is provided by a high dado of wood panelling which runs round the studio, a wood-strip floor and a plaster ceiling.

(6.3) Swansea, Studio 1

Swansea, Studio 1, has been fully described by Ward.[12] It is rectangular with a height only slightly less than the width. The volume is 36 000 ft^3 and the mean reverberation time is 1·3 sec. The studio was entirely rebuilt in 1952 after the war, having suffered extensive damage, and considerable latitude was possible in choosing the construction and acoustic treatment. Diffusion and bass absorption are obtained by the use of large numbers of line arrays of Helmholtz resonator absorbers in the form of rectangular hollow plaster castings applied to the wall surfaces. The ceiling is relieved with deep frames carrying porous absorbers covered with perforated plasterboard.

Diffusion and definition are good and the tone has a characteristic warmth.

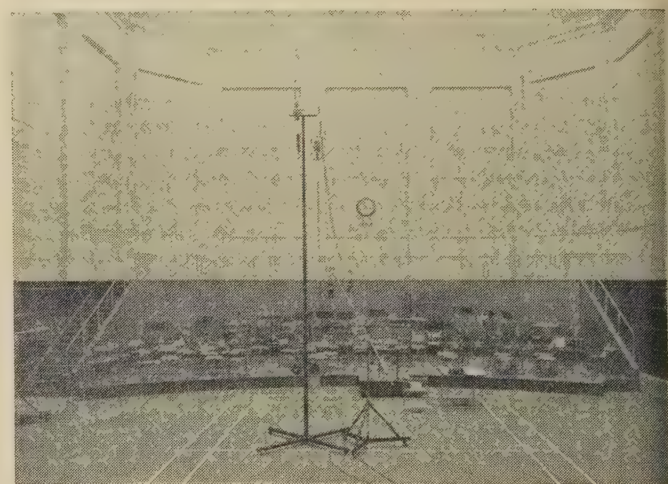


Fig. 11.—B.B.C. Concert Studio, Glasgow.

(6.4) Cardiff, Charles Street Studio

The Charles Street Studio, Cardiff, was formed from an existing church hall which had rendered brick walls pierced with large windows. It has a pitched timber roof and the space above the main roof trusses has been closed with a false ceiling. The remaining sloping soffits have been partly lined with glass wool covered with perforated hardboard. Generally, the wall and ceiling surfaces are fairly smooth with few scattering features. The reverberation time, averaging 1·3 sec, is rather long for the volume of 43 000 ft^3 and increases to 1·7 sec at 2 000 c/s. String tone is acceptable but definition is not very good, the bass in particular being insufficiently clear. Originally it was found necessary in playing to keep the tympani and brass down, but placing absorbing screens immediately behind these instruments largely cured the condition. It appears that the main defects are due to

rather low ceiling, lack of diffusion and too little absorption behind the orchestra.

(6.5) The Farringdon Hall, London

The Farringdon Hall, situated near Ludgate Circus in London, was put into use as a studio in 1951, almost without alteration. It has stone walls and a pitched timber roof rising to a height of 1 ft above the floor. Diffusion is improved by a balcony, the roof timbers and much decoration. The reverberation time, which averages 1·5 sec with a symphony orchestra, is about correct for the volume of 113 000 ft^3. It will accommodate the full B.B.C. Symphony Orchestra without giving the impression of being much too small, and the sense of spaciousness is comparable with that of the Maida Vale studio, which is approximately twice the volume. The definition and tone quality are both good, and the performance of this studio lends support to the impression that the height of a music studio should be little less than its width.

(7) SOUND DISTRIBUTION IN CONCERT HALLS

In many concert halls the distribution of sound in the auditorium is not uniform, and hence, in certain seats, the audience finds it difficult to hear all parts of the score. This defect is usually associated with standing waves, and therefore the sound is characterized by irregular frequency response. Furthermore, in many concert halls with deep balconies, listening conditions are often very poor underneath the balconies because the seats are badly screened.

It has been known for many years in connection with broadcasting studios that the best way to get a uniform sound distribution is to use scattering surfaces to produce diffusion. These surfaces take the form of irregularities, cylindrical forms being the most widely used in broadcasting. Observations have shown that, in all the good concert halls, uniform sound distribution is one of the very noticeable characteristics, and this always appears to be allied to the use of elaborate ornamentation, particularly on the ceiling.

It was the observation of good diffusion and richer tonal quality in studios and halls with rectangular coffering that led to an investigation by Somerville and Ward[13] which showed that a rectangular form is more efficient for diffusion than either cylindrical or triangular forms. This finding has since been substantiated by other workers[14, 15] but requires qualification. At high frequencies, at which the dimensions of the diffuser are large compared with the wavelength and reflection is specular, spherical and cylindrical surfaces are more effective than rectangular. It is, however, at medium and low frequencies that the greatest difficulty is experienced in obtaining sufficient diffusion because the dimensions of the diffusers are small compared with the wavelength. In these circumstances specular reflection is impossible and the important factor is the perturbation of the boundary surfaces. The maximum perturbation for given dimensions of the diffuser is produced by a rectangular form.

One of the principal reasons given[8] for the use of reflectors in modern concert halls is to increase the sound intensity in rear seats, particularly under balconies where screening is serious. There is no doubt that the direct sound is increased at the rear and under balconies by reflectors, but, because the sound strikes the audience before the sound energy can be built up, the reverberant sound is lacking. The high proportion of direct sound also appears to cause an impression of deadness even if the measured reverberation time is adequate.

The authors have recently measured the sound intensity at various points in a number of concert halls to try to find out whether reflectors are effective for the purpose for which they were designed. In each hall, measurements were made with a microphone placed close behind the conductor and with other microphones placed at positions to the rear of the auditorium. Direct measurements are possible at a rehearsal, but with an audience present this is very difficult, and therefore the procedure adopted was to make simultaneous recordings of the front microphone and one of the others. The microphones were previously calibrated. Subsequently the recordings were analysed with a level recorder to produce a chart of the type shown in Fig. 12. Analysis gave the results shown in Figs. 13, 14, 15, 16 and 17. It will be seen that, as regards sound distribution, the Royal Festival Hall is the least satisfactory, and that there

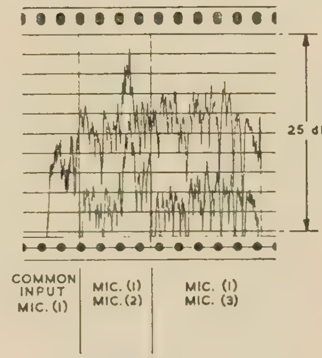


Fig. 12.—Typical level-recorder traces, showing the differences between the levels in three microphone positions.

Writing speed: 25 dB/sec.
Length of extract: 55 sec.

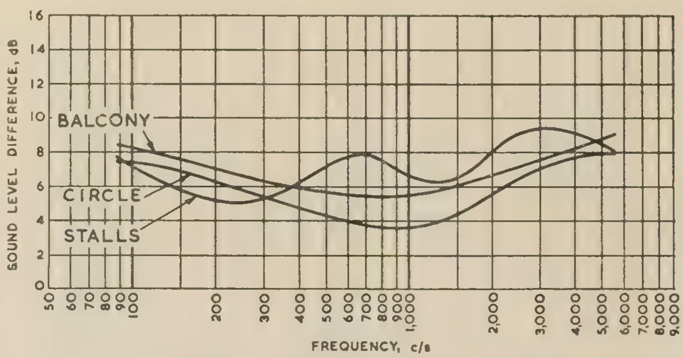


Fig. 13.—Sound-level differences between front and rear seats, Free Trade Hall, Manchester.

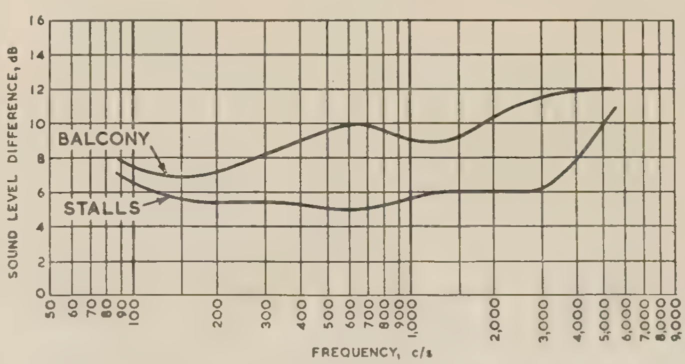


Fig. 14.—Sound-level differences between front and rear seats, Liverpool Philharmonic Hall.

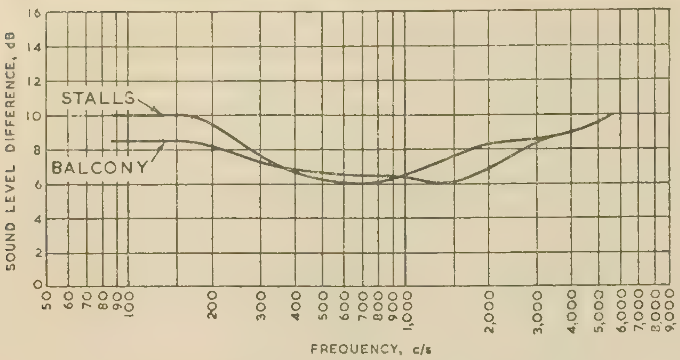


Fig. 15.—Sound-level differences between front and rear seats, St. Andrew's Hall, Glasgow.

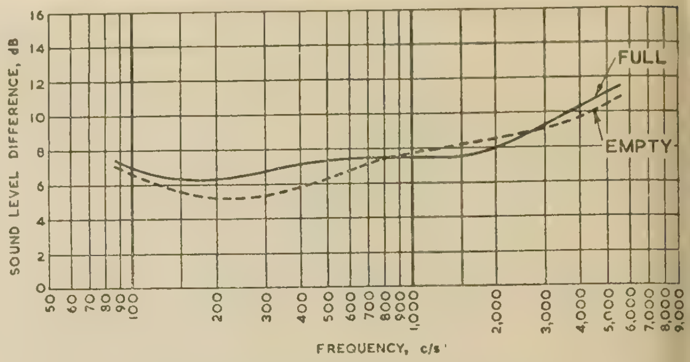


Fig. 19.—Effect of audience on sound-level difference between front and back, Liverpool Philharmonic Hall.

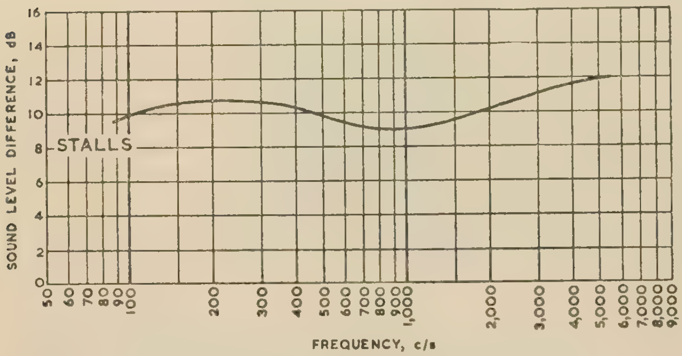


Fig. 16.—Sound-level differences between front and rear seats, Usher Hall, Edinburgh.

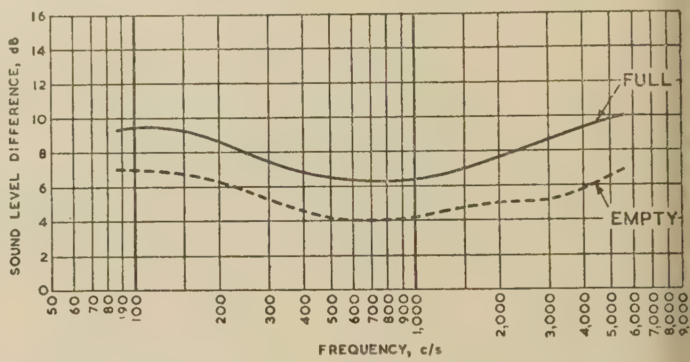


Fig. 20.—Effect of audience on sound-level difference between front and back, St. Andrew's Hall, Glasgow.

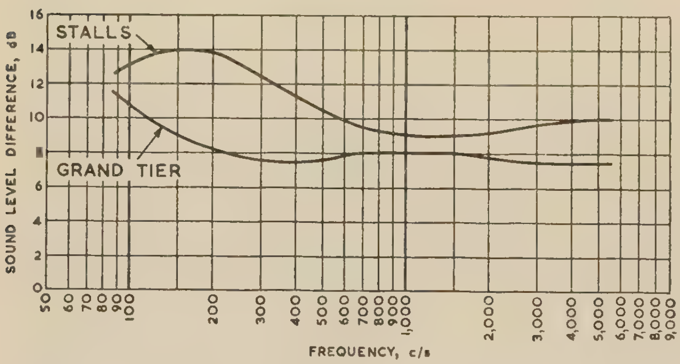


Fig. 17.—Sound-level differences between front and rear seats, Royal Festival Hall, London.

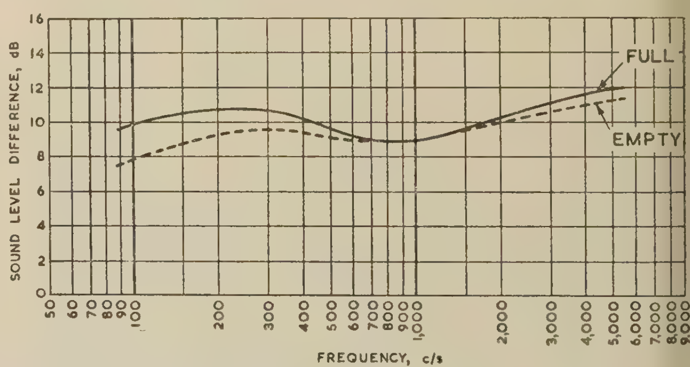


Fig. 21.—Effect of audience on sound-level difference between front and back, Usher Hall, Edinburgh.

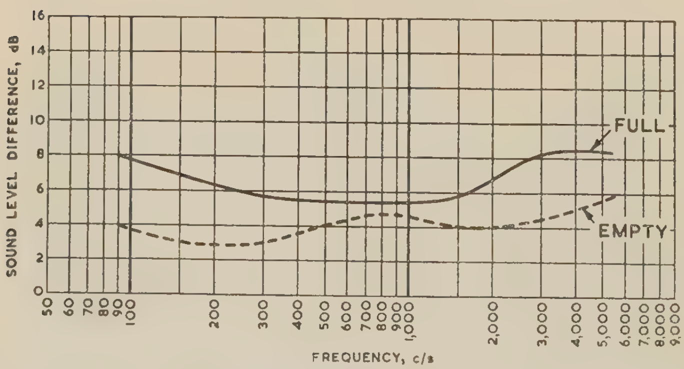


Fig. 18.—Effect of audience on sound-level difference between front and back, Free Trade Hall, Manchester.

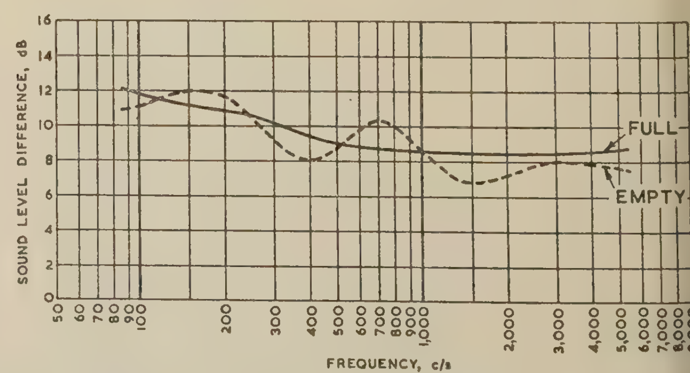


Fig. 22.—Effect of audience on sound-level difference between front and back, Royal Festival Hall, London.

not much to choose between any of the others. Although the Royal Festival Hall has indeed the largest volume of the halls investigated, its length of 144 ft from centre stage to the extreme rear is no greater than that of St. Andrew's Hall (144 ft) and slightly less than that of the Liverpool Philharmonic Hall (148 ft). The other two halls are slightly shorter. The effect of the audience is shown in Figs. 18, 19, 20, 21 and 22. As would be expected, the differences due to the audience are most marked in the case of St. Andrew's Hall, where the seating is very old and somewhat austere.* The results of these experiments fail to substantiate the claims made for reflectors. Bearing in mind that the lack of diffusion in halls with reflectors results in hard tonal quality and also produces the effect of masking by the powerful instruments, it is clear that the use of reflectors is not to be recommended.

(8) AURAL ASSESSMENT

Although much experience has been gained in the design of large auditoria and broadcasting studios, the only method of measurement on which there has been reasonable agreement is that of reverberation time. Using reverberation time as the objective criterion, there have been considerable variations reported by many workers, and in 1936 Kirke and Howe[16] published details of an experiment in which the B.B.C. constructed two studios identical in volume and reverberation time, but with different interior treatments. The fact that these studios differed considerably in acoustic properties as judged subjectively demonstrates convincingly that reverberation time may be inadequate as a criterion. Consequently, workers in the field of acoustics have to rely very much on the opinions of critical listeners who are able to assess acoustic characteristics without objective assistance.

Skilled observers can detect the effects of poor diffusion and can hear the harsh tone produced by sound decays which are not smooth. Much can be learnt about an auditorium by closing one ear and moving around. If diffusion is good, there will be little variation in the sound field and it should be possible to hear all the parts in an orchestra. However, if the diffusion is not good and therefore pronounced standing waves exist, considerable variation will be observed. With monaural hearing the normal directional properties are inhibited and acoustic faults are emphasized. This, of course, is the condition in a broadcasting studio or concert hall when a microphone is being used, and is the reason for the difficulty in obtaining acoustics good enough for broadcasting.

(9) SUBJECTIVE COMPARISONS AND CORRELATION WITH OBJECTIVE MEASUREMENTS

(9.1) British Investigations

In 1952, Parkin, Scholes and Derbyshire[4] described a subjective investigation on the acoustic properties of a number of British concert halls by means of a questionnaire sent to well-known people in the world of music. This investigation showed that the subjects who knew all the concert halls considered that the Liverpool Philharmonic Hall was better than St. Andrew's Hall, Glasgow, and Usher Hall, Edinburgh, less good than either. At the time of this investigation the Royal Festival Hall had not been built. At a later date a B.B.C. investigation was carried out[17] using recordings of the same work in a number of concert halls. The preference in this case was for St. Andrew's Hall, with the Royal Festival Hall second and Usher Hall third; it was not possible to include the Liverpool Philharmonic Hall. In this investigation the only group of subjects giving significantly concordant answers were specialists skilled in balancing or criticizing musical programmes. Engineers, performing musicians and the general public were less concordant. In 1953, Parkin, Allen, Purkis and Scholes, in describing the acoustics of the Royal Festival Hall,[8] expressed a similar opinion. One of the conclusions of the B.B.C. investigation is of interest:

> It is disappointing, although not unexpected, that the general public is quite unable to produce significant results, for the skilled listeners employed in this investigation, in common with the subjects used in Parkin's investigation, must necessarily form a small percentage of any concert audience.

(9.2) German Investigations

Recently W. Kuhl[18] has endeavoured to find the preferred reverberation time for broadcast programmes by making recordings in many concert halls and studios in Germany. He has produced some surprising results in that the preferred reverberation time for modern music is found to be as short as 1·5 sec. B.B.C. experience during the last ten years is not in agreement with this view, which conflicts also with all previously published estimates of optimum reverberation times for large concert halls. Provided that the usual criteria for good design have been observed, it is normal practice to perform modern and romantic music in studios with reverberation times lying between 1·7 and 1·9 sec, and no difficulty is experienced either in performance or microphone placing.

In making recordings for the comparison of concert halls certain precautions must be observed. These have been described in a paper by one of the authors.[17] The same type of microphone should be used for each performance, and the microphone positioning should not be left to the normal broadcasting personnel. The professional operator will endeavour to obtain the best possible result, as he should for a broadcast, but in the process he may produce a balance which conceals acoustic faults of the hall. In fact, the result will be coloured by his own subjective judgment. If the acoustics present difficulty the inevitable procedure is to place the microphone, or microphones, close to the orchestra, thereby reducing the effective reverberation time. Therefore in the B.B.C. experiments every precaution was taken to ensure that the recorded result was characteristic of the hall. The same three observers co-operated in the microphone placing in each hall and they were not the operators accustomed to the halls. Their terms of reference were to produce a balance characteristic of the concert hall, which is not necessarily coincident with the best broadcast quality. It is necessary to standardize the monitoring conditions, which is done by using an acoustically treated van, and listening in every case at the same loudness, established by measurement, since a level change would alter the relationship between the loudness of different parts of the musical scale. Similarly, in reproduction, standard listening conditions are essential as regards acoustics, equipment and listening level, which should be that used in the original monitoring.

During the subjective tests it was found that the opinions of subjects could be altered completely by changing the loudness at which they listened. This was done experimentally and the results were published in the previous paper. The final evaluation was carried out only with opinions expressed when listening at the same loudness as had been used for the original monitoring. Since the position of the microphone depends on the loudspeaker used for monitoring,[19] the same microphone and loudspeaker were used throughout.

In Kuhl's investigation, the recordings used for the subjective assessments were taken from a microphone placed by an experienced broadcasting engineer to give the best reproduced results, the only restriction being a minimum distance of 5 metres from the nearest instrument. His findings cannot therefore be held to apply to concert halls used as such. It is not stated

* Recently this hall has been reseated and redecorated, but information on the effects is not yet available.

whether the necessary precautions in connection with the listening tests, as outlined above, were observed.

A further experiment has been conducted by Reichardt, Kohlsdorf and Mutscher,[20] using a method of assessment of the subjective results similar to that of Kuhl. The music consisted of string music, light opera and film scores recorded in the rather dead Unter den Linden Opera House, Berlin. Reverberation was added by the use of an 'echo room', and the effective reverberation times preferred by the subjects varied from 1·2 to 1·6 sec. In view of the poor quality of the reverberant sound in an 'echo room' of only 53 m^3, however, it is indeed surprising that the subjects would tolerate the addition of even moderate proportions of reverberation. The experiment, therefore, gave no information about optimum reverberation time under good concert-hall conditions.

In Fig. 2 is drawn the B.B.C.'s optimum reverberation curve, based on past experience; deviations upward from the indicated values are possible if other aspects of the acoustics are good, but times very much shorter are not generally associated with good orchestral quality.

(9.3) An Empirical Acoustic Criterion

One of the authors has already published details of an objective criterion[21] which gives reasonable agreement with subjective

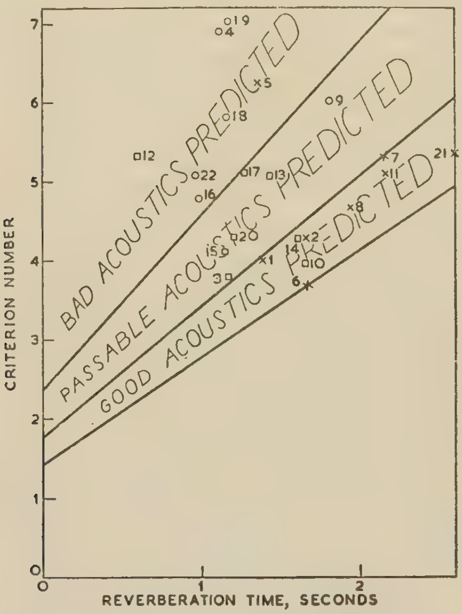


Fig. 23.—Empirical acoustic criterion.

Subjective grading:
× Good acoustics.
□ Passable acoustics.
○ Bad acoustics.

Reverberation times with no orchestra or audience:

1. London, Broadcasting House, Concert Hall.
2. Manchester, Milton Hall.
3. Belfast, Studio 1.
4. Glasgow, Studio 2.
5. Glasgow, Studio 1.
6. London, Maida Vale, Studio 1.
7. Bristol, Colston Hall.
8. Manchester, Free Trade Hall.
9. London, Royal Festival Hall.
10. Liverpool, Philharmonic Hall.
11. Edinburgh, Usher Hall.
12. London, Maida Vale, Studio 5.
13. London, Camden Theatre.
14. London, Criterion Studio.
15. London, Paris Studio.
16. Birmingham, Studio 4.
17. Birmingham, Vestry Hall.
18. Belfast, Studio 8.
19. Manchester, Studio 1.
20. Edinburgh, Studio 1.
21. Glasgow, St. Andrew's Hall.
22. Bristol, Studio 1.

assessments. Further work has since been carried out to extend the criterion by obtaining the opinions of a large number of subjects, and a paper was read at the Second International Congress on Acoustics at Cambridge, Massachusetts, in June, 1956. The paper will be published in *Acustica* in March, 1957.

The criterion number is given by the equation

$$X' = (D + 0{\cdot}332R)/(1{\cdot}64 + 1{\cdot}542T_m)$$

where D = Parameter measuring the irregularity of decays.
R = Parameter measuring the irregularity of the reverberation-time/frequency characteristic.
T_m = Mean reverberation time.

The details of the derivation of these parameters are described in the first paper.[21] The results are plotted in Fig. 23. In this Figure the subjective assessment of the various enclosures is indicated, and it will be seen that agreement between the objective criterion and the subjective assessment is good.

(10) CONCLUSIONS

In the paper an attempt has been made to crystallize the results of subjective and objective experiment and of observation over the last ten years or so. The subject under investigation is purely aesthetic and therefore must begin and end with human aesthetic judgments. The difficulty of obtaining concordant judgments has already been remarked upon, and most previous investigations have been limited in their validity by this difficulty. The B.B.C. is fortunate in having available a large number of people who are professionally employed in capacities requiring accurate auditory memories (a rather uncommon gift) and consistent aesthetic judgment. Much use has been made of their opinions in comparing concert halls and studios, and the judgments of music critics, particularly in connection with the new concert halls, have been followed over a long period.

The modern type of concert hall, characterized by a fan-shaped plan with a reflector over the stage, splays at the side and a concave ceiling, all designed to provide strong first reflections to reinforce the direct sound, has serious disadvantages from the aesthetic point of view. The tonal quality is generally unsatisfactory, and although there is a first impression of 'clear' definition owing to the high level of direct relative to reverberant sound, this is largely illusory as inner parts of the score tend to disappear in loud passages.

In most but not all the newer concert halls reviewed, the harsh tone is also associated with a short reverberation time. The reverberation time of most new halls is less than is desirable and in some halls less than the designers intended. This should be taken into consideration in future designs.

The use of reflecting surfaces to project sound away from the orchestra introduces bass masking even if the low-frequency reverberation time of the hall is not excessive. The effect of this is noticed in most modern halls as a deterioration of definition in loud passages.

Quite apart from these disadvantages of the 'sound reinforcing' type of enclosure, it has been shown also that the use of reflectors and similar devices has not resulted in any increase in the sound level at the back of the hall as compared with existing halls of traditional shape. This failure to achieve their original object is explained by the rapid absorption of direct and once-reflected sound by the audience areas and consequent reduction of the reverberant-energy contribution.

It will be clear from the progress of modern design from the Salle Pleyel onwards that concert-hall designers have found it necessary to discard, one by one, the principles adopted in its design, and the authors are convinced that no justification exists for further experiments of this kind.

For future concert halls, architects must revert to forms of design which are modern realizations of the features which combined to give the fine tonal quality and clear definition of the best of the traditional halls.

These features are the provision of adequate scattering, absence of deliberate reinforcements of the direct sound, ample height and an adequate reverberation time.

(11) ACKNOWLEDGMENTS

The authors wish to thank the Chief Engineer of the British Broadcasting Corporation for permission to publish the paper. They also wish to acknowledge the assistance of their colleagues in the B.B.C. Research Department, and in particular that of Mr. F. L. Ward. Thanks are also due to the managements of all the concert halls mentioned in the paper.

(12) REFERENCES

(1) SOMERVILLE, T.: 'A Comparison of the Acoustics of the Philharmonic Hall, Liverpool, and St. Andrew's Grand Hall, Glasgow', *B.B.C. Quarterly*, 1949, **4**, p. 41.

(2) *The Times*, 20th December, 1954.

(3) PARKIN, P. H.: 'Acoustics of the Royal Festival Hall', *Nature*, 1951, **168**, p. 264.

(4) PARKIN, P. H., SCHOLES, W. E., and DERBYSHIRE, A. G.: 'The Reverberation Times of Ten British Concert Halls', *Acustica*, 1952, **2**, p. 97.

(5) BAGENAL, H., and WOOD, A.: 'Planning for Good Acoustics' (Methuen, 1931), p. 268.

(6) ANDRADE, E. N. DA C.: 'The Salle Pleyel', *Nature*, 1932, **130**, p. 332.

(7) ALLEN, W. A., and PARKIN, P. H.: 'Acoustics and Sound Exclusion', *Architectural Review* (London), 1951, **109**, p. 377.

(8) PARKIN, P. H., ALLEN, W. A., PURKIS, H. J., and SCHOLES, W. E.: 'The Acoustics of the Royal Festival Hall', *Acustica*, 1953, **3**, p. 1.

(9) SHAWE-TAYLOR, D.: 'The Festival Hall', *New Statesman and Nation*, 27th October, 1951.

(10) ALLEN, W. A.: 'Acoustics in the New Concert Halls', *Journal of the Royal Institute of British Architects*, 1951, **59**, p. 39.

(11) GILFORD, C. L. S.: 'Membrane Sound Absorbers and their Application to Broadcasting Studios', *B.B.C. Quarterly*, 1952–53, **7**, p. 246.

(12) WARD, F. L.: 'Helmholtz Resonators as Acoustic Treatment in the New Swansea Studios', *ibid.*, p. 174.

(13) SOMERVILLE, T., and WARD, F. L.: 'Investigation of Sound Diffusion in Rooms by Means of a Model', *Acustica*, 1951, **1**, p. 48.

(14) BRÜEL, P. V.: 'Automatic Recording of Frequency Irregularity in Rooms', *ibid.*, 1954, **4**, p. 21.

(15) LAMORAL, R., and TREMBASKY, R.: 'Investigation of the Acoustics of Large Halls using Models', *L'Onde Electrique*, 1953, **33**, p. 570.

(16) KIRKE, H. L., and HOWE, A. B.: 'The Acoustical Design of Broadcasting Studios', *Journal I.E.E.*, 1936, **78**, p. 404.

(17) SOMERVILLE, T.: 'Subjective Comparison of Concert Halls', *B.B.C. Quarterly*, 1953, **8**, p. 125.

(18) KUHL, W.: 'Nachhallzeit Grosser Musikstudios', *Acustica*, 1954, **4**, p. 618.

(19) SOMERVILLE, T.: 'Monitoring Sound Broadcast Programmes', *Wireless World*, May, 1956, p. 228.

(20) REICHARDT, W., KOHLSDORF, E., and MUTSCHER, H.: 'Die Optimale Nachhallzeit für Studioräume', *Hochfrequenztechnik und Elektroakustik*, 1955, **64**, p. 18.

(21) SOMERVILLE, T.: 'An Empirical Acoustic Criterion', *Acustica*, 1953, **3**, p. 365.

621.396.621

Paper No. 2151 R
Aug. 1956

FREQUENCY DIVERSITY IN THE RECEPTION OF SELECTIVELY FADING BINARY FREQUENCY-MODULATED SIGNALS

With special reference to Long-Distance Radiotelegraphy

By J. W. ALLNATT, B.Sc.(Eng.), Associate Member, E. D. J. JONES and H. B. LAW, B.Sc.Tech., Associate Member.

(*The paper was first received* 27*th February, and in revised form* 14*th May*, 1956. *It was published in August*, 1956, *and was read before the* RADIO AND TELECOMMUNICATION SECTION 14*th November*, 1956.)

SUMMARY

In current long-distance h.f. radio practice frequency-shift-telegraphy signals are usually demodulated by the use of limiters and discriminators on conventional frequency-modulation lines. This method of demodulation is liable to fail whenever the signal fades on either the mark frequency or the space frequency. This is wasteful. The full message is available on each frequency and failure need not occur unless signals on both fade together; if the fading is frequency-selective, as it often is, simultaneous fading may be comparatively rare. Thus there is the possibility of frequency diversity.

Theoretical analysis of the case of Rayleigh-fading signals disturbed by white Gaussian noise, the fading being slow relative to the speed of signalling, shows that substantial advantages may be derived from frequency diversity. Also, the analysis leads to a mathematical specification for an ideal receiver. An experimental equipment has been made, which, while avoiding the complexity that full accord with the specification would involve, complies with it in several respects. The most important feature is that the received signals are assessed in terms of the expected amplitudes of mark and space signals, derived from observation of earlier elements; that is to say, the absence of a mark signal, which would be strong if it were present, is taken as a strong indication of space. The effective bandwidth is reduced almost to the limit by the use of linear-build-up band-pass filters to respond to the mark and space frequencies.

Laboratory tests on the experimental unit gave results in good agreement with theory. At the optimum signalling speed the noisy-signal performance was only about 3 dB short of the ideal. On signals from Australia the experimental unit gave a useful improvement compared with a high-grade receiver of the conventional type, although not so large as in the laboratory tests. The difference is attributed to the occurrence on the practical radio channel of disturbances, such as interference and atmospheric crashes, other than white Gaussian noise. Further investigations are needed.

The new technique may have other applications.

LIST OF PRINCIPAL SYMBOLS

b = Characteristic signal/noise energy ratio, giving error rate $1/2\varepsilon$ when the exponential error-liability versus signal/noise-ratio characteristic applies.
m = Undistorted mark signal waveform.
N_0 = Noise power per unit bandwidth.
p, p_a, p_b, p_m, p_s = Instantaneous signal power. Suffixes represent propagation paths a and b, mark signal and space signal.
P = Probability.
P_e = Probability of error (steady signals).
p = Probability density.
s = Undistorted space signal waveform.
T = Duration of one telegraph element.
w = Signal energy for a particular telegraph element.
W = Mean signal energy per element.
$w_a w_b$, $W_a W_b$ = Energies received by propagation paths a and b.
w_d = Diversity component of received energy.
w_e = Effective signal energy per element.
w_v = Non-diversity component of received energy.
W_1 = Mean energy per element received on aerial 1.
y = Received waveform (signal plus noise).
Δf = Frequency difference between mark and space.
δf = Receiver tuning error.
η = Efficiency of utilization of received energy.
θ = A time-variable phase difference.
ν = Complex frequency.
τ = Path-time-delay difference in two-path propagation.
ϕ = Phase difference corresponding to τ and Δf.
ω = Angular frequency.

The authors are at the Post Office Research Station.

(1) INTRODUCTION

Long-distance radio-telegraphy in the 4–30 Mc/s frequency band is much handicapped by fading, noise, atmospherics and interference. To minimize the effects of these disturbances, important point-to-point channels are normally equipped with highly directional diversity aerials and elaborate receivers, but, even so, they are at times so badly affected as to be unserviceable.[1] It is thus of high practical importance that there should be no unnecessary loss of performance in receiving equipment. To establish this condition requires a study of the information available at the receiver and of the manner in which receivers fail when operating conditions deteriorate.

It is a common and increasing practice to use frequency-shift telegraphy (f.s.t.) on long-distance radiotelegraph channels, the signals usually being demodulated in the receiver by some kind of limiter/frequency-discriminator arrangement. Such a combination is well able to absorb the variations of signal amplitude arising from fading, and, so long as a reasonable signal/noise ratio is maintained at the limiter input, post-discriminator filtering gives the usual advantage of frequency modulation, and telegraph distortion is reduced relative to that of an amplitude-modulation system of the same transmitted power. If the input signal/noise ratio falls to the neighbourhood of unity, the improvement by f.m. operation is lost and any telegraph signal element received under these conditions is liable to be wrongly interpreted; i.e. in a system using a binary code a mark signal may be taken for a space, and vice versa. The performance of receivers of the f.m. type when subject to disturbance by white Gaussian noise has been studied,[2] and, provided that the fading of the signal is not frequency-selective, performances within a few decibels of the ideal can be obtained with good equipment.

Long-distance h.f. radio channels are, however, very prone to

selective fading. This arises under conditions of multi-path propagation whereby significant signal components reach the receiver by two or more different paths having different propagation times; path-time-delay spreads of the order of 1 millisec are commonly observed. The signals arriving by the different routes fade independently, so that the effective propagation time from transmitter to receiver varies as first one and then another route gains ascendancy; this causes a form of telegraph distortion known as varying bias, which may be aggravated, especially if a small frequency shift is used, by distortion of the waveform of the discriminator output during transitions between mark and space, owing to beating between differently delayed signal components. These troubles are rarely serious with the frequency shifts and telegraph speeds used in practice on point-to-point services, which are typically of the order of 500 c/s and 100 bauds. It may be concluded that multi-path propagation cannot improve a radiotelegraph channel on which the normal f.m. technique of reception is used and that it may degrade its performance, whereas with an ideal receiver it can give an improvement.

In normal messages with the binary telegraph codes generally used in synchronous telegraph systems, only a small number of elements of a given type can be produced in uninterrupted succession, i.e. without elements of the other kind occurring. This being so and fading being very slow relative to the speed of signalling, it follows that, if the mark and space frequencies fade differently owing to multi-path propagation, it is likely that a weak element or group of elements of one kind will be preceded and followed by comparatively strongly received elements of the other kind. If, for example, two strongly-received mark-signal elements are separated by a weak and indeterminate element, this latter can be identified with certainty as a space element, for had it been a mark element it would have been strongly received; uncertainty need only arise on occasions, which may be comparatively rare, when both mark and space are badly received. Multi-path propagation is therefore potentially useful, for there is a possible frequency-diversity advantage. This advantage is lost in the normal f.m. method of reception, in which the limiter destroys the vital information about signal amplitude. These considerations suggest that a quantitative investigation of the reduction in error liability to be gained by the application of frequency diversity is desirable.

(2) PRELIMINARY CONSIDERATIONS

(2.1) Performance of Ideal Receiver

For good performance under adverse conditions of operation a receiver must make proper use of all the available information, including ideally the waveforms of the various kinds of signal that might be received in the absence of disturbances and the probabilities of their being sent, as well as the waveforms of the disturbed signals actually received. The interpretation of signals that are affected by Gaussian noise or a similar random disturbance must be subject to some uncertainty, and Woodward and Davies have postulated[3] an ideal receiver, the output of which is a measure of the probability that a signal exists. For practical purposes of binary-code telegraphy a receiver must interpret each signal element as a mark or a space, and an ideal telegraph receiver may be defined as one that does this with the minimum possible probability of error. An indication of the advantage to be gained by frequency diversity may be obtained by considering the performance of such a receiver under various conditions of selective fading. It is assumed that mark and space are equally likely to be sent.

The error liability after regeneration of signals received by an ideal diversity telegraph receiver fed with Rayleigh-fading signals plus white Gaussian noise has been determined.[4] The best that can be hoped for in frequency diversity is that the fadings of the mark and space frequencies be uncorrelated; in these circumstances a single-aerial frequency-diversity receiver is equivalent to the two-diversity-branch case described in Reference 4, and the two-aerial diversity receiver is equivalent to the four-diversity-branch case. These two cases give the curves marked '0' in Fig. 1. The signal/noise ratio is expressed in terms of the noise

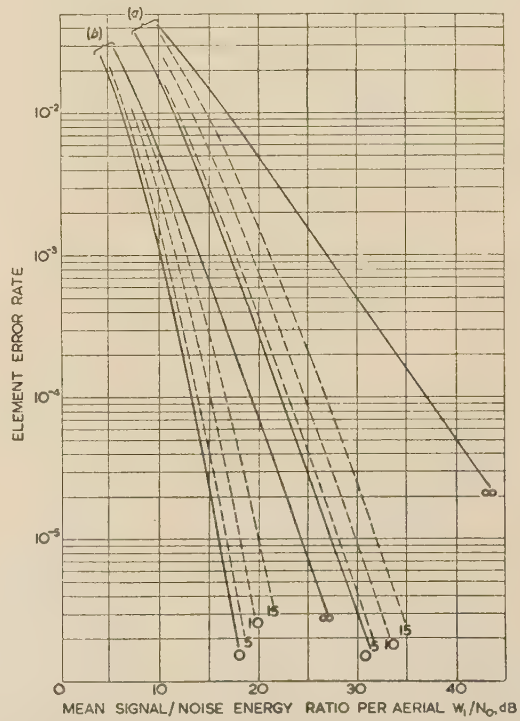


Fig. 1.—Error liability in ideal diversity reception of Rayleigh-fading signals under two-path propagation conditions with ideal path-time-delay difference.

(a) Single-aerial case.
(b) Two-aerial case.
Parameters on curves give relative activities of paths in decibels.

power per unit bandwidth, assumed to be the same for the two aerials and the mean signal energy per element per aerial, also assumed to be the same for the two aerials. It is to be noted that the energy is given on a per-aerial basis rather than on the per-branch basis used previously. If the fading is flat the frequency-diversity is lost and the performances of the single-aerial and two-aerial receivers revert respectively to the one- and two-branch cases of Reference 4, with the difference, however, that an advantage of 3 dB arises from the fact that the signals derived from the mark and space frequencies are coherent, whereas the noises are not. The curves marked '∞' in Fig. 1 give the performance under flat-fading conditions.

Some idea of the performance of the ideal receiver under conditions between the extremes just discussed has been obtained from a study of the two-path-propagation case (see Section 9.1). This case was chosen for its simplicity and for the reason that it can be simulated by existing laboratory equipment. The optimum condition of uncorrelated fading on the mark and space frequencies is obtained if the two paths are equally active and the path-time-delay difference, τ, is half the reciprocal of the difference, Δf, between mark and space frequencies; the performance deteriorates if the conditions depart from the optimum either in respect of relative path activity or path-time-delay difference. The curves in Fig. 1 show the performance of the ideal receiver

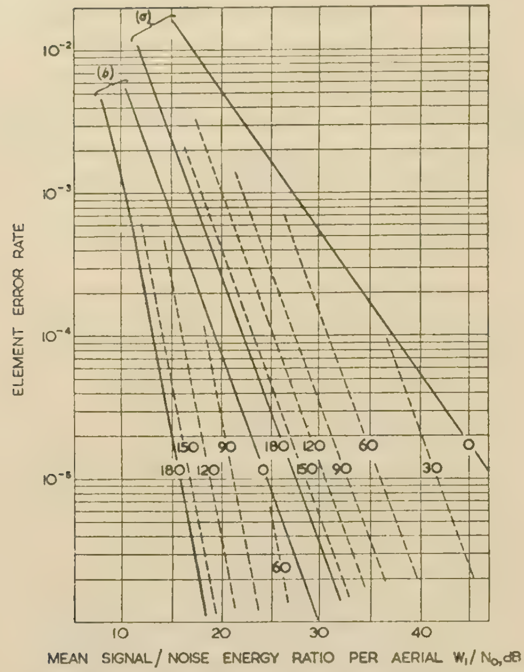


Fig. 2.—Two-path propagation; paths equally active. Effect of variation of path-time-delay difference.

(a) Single-aerial case.
(b) Two-aerial case.

Parameters on curves are values of ϕ = (p.t.d. difference) × (frequency shift) × 360 in degrees.

for various ratios of the powers in the two paths, with optimum path-time-delay difference. The case of non-ideal path-time-delay difference is less amenable to calculation. The curves in Fig. 2 relate to this case, the path activities being assumed equal, and they represent lower limits of performance for various values of a phase angle ϕ given by

$$\phi = 360\tau\Delta f \text{ degrees} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

The error liability for given values of ϕ and mean signal/noise energy ratio per aerial will not exceed that given by the relevant ϕ-curve or by the ($\phi = 0$)-curve, whichever is the smaller.

The curves in Figs. 1 and 2 show that under favourable conditions the advantage to be gained by frequency diversity may be substantial. The two-path-propagation case shows appreciable advantage over a wide range of conditions, and it is to be expected that any increase in the number of propagation paths contributing significantly to the received signal power will in general tend to improve the frequency diversity towards the limit corresponding to zero correlation. Clearly, therefore, frequency diversity is potentially useful.

(2.2) Design Implications Arising from the Mode of Action of the Ideal Receiver

Reverting to the idea of existence probability mentioned in Section 2.1, an ideal receiver for binary-code telegraphy according to the definition given will identify a signal as a mark if the existence probability of mark exceeds 0·5. For an ideal receiver with q diversity branches, of which p are devoted to the recognition of mark signals and $q - p$ to the recognition of space signals, mark and space being equally likely to be sent, previous work has shown[4] that for the existence probability of mark to exceed 0·5 we have

$$\int_0^T \left[\sum_{u=1}^{p} (2y_{mu}m_u - m_u^2) + \sum_{v=1}^{q-p} (s_v^2 - 2y_{sv}s_v) \right] dt > 0 \quad . \quad (2)$$

where y_{mu} and y_{sv} are the waveforms of signals plus noise in the uth mark and vth space branches respectively, and m_u and s_v are the waveforms that would be received in the respective branches if noise were absent and mark or space respectively were sent. The noise power per unit bandwidth is assumed to be the same in all diversity branches. The undisturbed waveforms m_u and s_v are assumed to be completely known in all respects, including carrier phase. The time T is the duration of one signal element, zero time being the start of the element being interpreted. In a practical system p and $q - p$ would, of course, be the same and equal to the number of diversity aerials.

It would probably be practicable to design a correlation receiver to function precisely on the basis of expression (2), but it would not be simple, for some device would be necessary in each diversity branch to keep track of the carrier phase, which would fluctuate continuously with the fading. Furthermore, phase-continuity requirements would impose restrictions at the transmitting end. It was therefore decided that, in a first attempt to realize frequency diversity, carrier phase should be ignored in the interests of simplicity. Although expression (2) is thus rejected as a complete basis for design, it nevertheless gives some important indications. In the first place, it shows that the conbributions of the various diversity branches should be added together, rather than the best one selected; the effect of this on performance is discussed elsewhere.[2] Secondly, the individual contributions should be weighted according to the time integrals, over an element, of the squares of the signal voltages, or, in other words, according to the signal energies. These two points are in accordance with the conclusions reached by Kahn.[5] Finally—and this is most important—the contribution of any branch can be of either sign, so that a mark branch can give a strong indication of a space condition; it is by the recognition of absence of mark and its interpretation as presence of space (and similarly the recognition of absence of space) that the frequency-diversity advantage is obtained.

The bandwidth of the ideal receiver must be the optimum for the speed of signalling, and this is much smaller than the bandwidths employed in conventional receivers. This necessitates more accurate control of transmitter and receiver frequencies. Consideration of expression (2) shows that, if there is a frequency error δf between the signals received and the local signals with which they are correlated, the efficiency of utilization, η, of the received signal energy is given by

$$\eta = \frac{\sin(\pi\delta f T)}{\pi\delta f T} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

i.e.

$$\eta \simeq 1 - \tfrac{1}{6}(\pi\delta f T)^2 \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Thus for an efficiency higher than 90%, corresponding to a maximum loss of about 0·5 dB, we have

$$\delta f < \frac{1}{4T} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

A speed of signalling of 100 bauds therefore requires the error in frequency alignment of the transmitting and receiving equipments relative to each other to be less than 25 c/s. This is to be considered in relation to carrier frequencies in the 10–20 Mc/s region, where the frequency tolerance permitted by international regulations is 30 parts in 10^6; although the stability required is far greater than that called for by the regulations, it is, nevertheless, readily attainable in simple equipment.

A simple way of obtaining a measure of the mark-frequency energy in a signal element would be to apply the signal to a filter responsive to the mark frequency and having a build-up that is linear and that occupies a time equal to the duration of one element. This could be done by means of a high-Q resonator

ie oscillations in which are quenched after their magnitude has ›een determined at the end of each element, as described by oelz,[6] but the same result may be achieved, without the need ›r quenching, by a suitably designed filter network. If signals ere fed to such a filter and thence to a linear detector the sult, assuming a reasonable signal/(noise-plus-interference) tio, would be a voltage varying rapidly, according to the keying, etween relatively high values corresponding to marks and latively low values corresponding to spaces; the high and low alues themselves would vary comparatively slowly according to ie fadings of signal and interference. Evidently, the detector utput could be interpreted in relation to a 'judgment level' oout half-way between the high and low levels, any signal ement being interpreted as a mark or a space according as this etector output lay above or below the judgment level existing

diversity advantage. It has not been necessary to build a complete receiver to do this, because it is only the method of demodulation that is in question; the early stages of good-class communication receivers are suitable for feeding the new type of demodulation unit. The experimental equipment has therefore been designed to be fed from the final i.f. amplifiers of a high-grade telegraph receiver which is extensively used at British Post Office receiving stations. In this way the experimental equipment benefits from the selective channel filters incorporated in the receiver. Moreover, since the interconnection of the conventional and the experimental equipments is by means of high-impedance teeing pads, which leave the normal operation of the standard equipment unaffected, comparisons between the new method of demodulation and the limiter/discriminator arrangement of the conventional equipment are easy to make. The

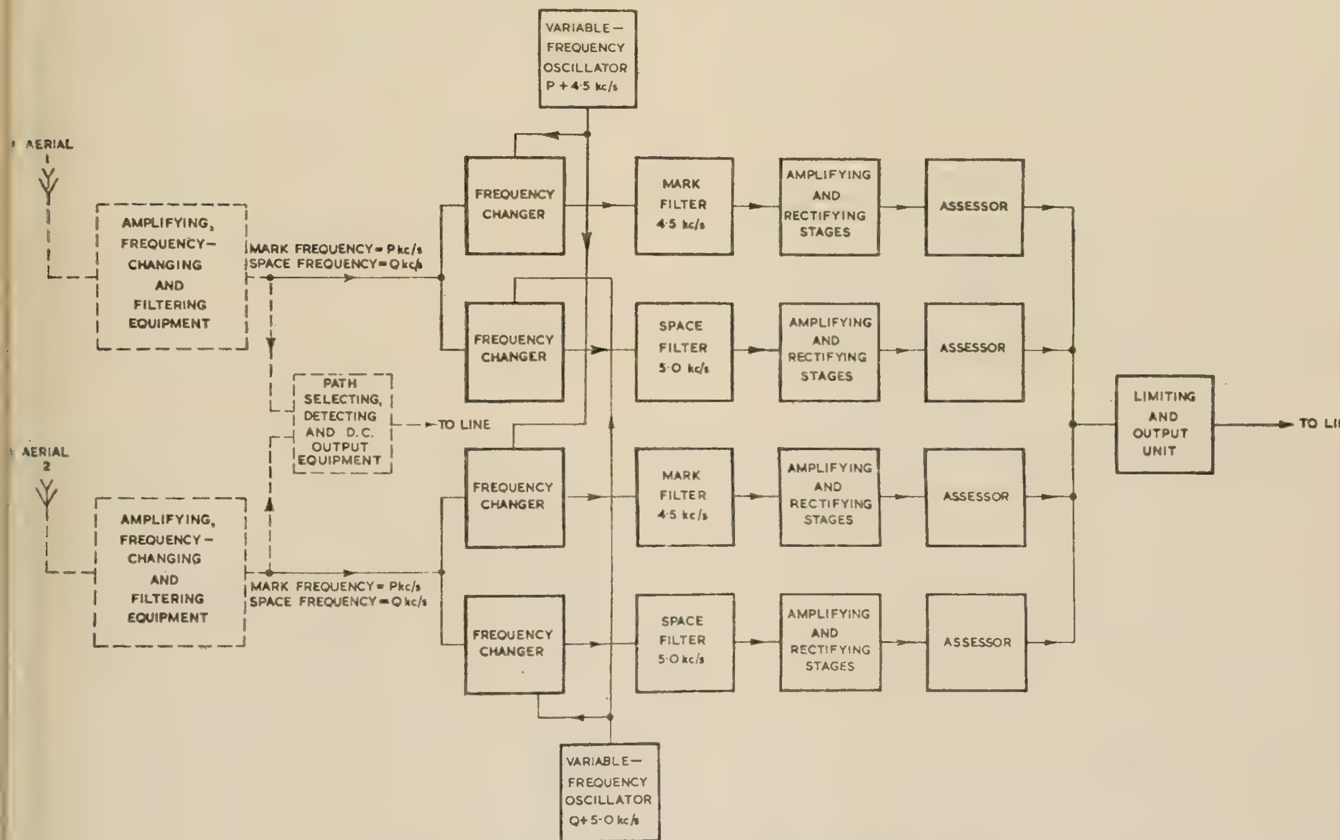


Fig. 3.—Demodulation unit as connected to existing receiver.

---- Parts of existing receiver.
—— New demodulation unit.

it the time, and the strength of the interpretation being indicated y the magnitude of the difference between the detector output nd the judgment level. The two-aerial receiver would have our diversity branches on the lines indicated, two for mark and wo for space, the outputs having to be suitably combined. An xperimental equipment of this kind has been constructed.

(3) EXPERIMENTAL DEMODULATION UNIT

(3.1) General

In building the experimental equipment the primary object has een to facilitate and expedite laboratory and field trials intended o test the practicability of obtaining a worth-while frequency-

block schematic (Fig. 3) shows the experimental demodulation unit working under these conditions.

The mark and space frequencies of 4·5 and 5·0 kc/s, respectively, are derived from the two i.f. signals of about 100 kc/s by mixing them with the outputs of two variable-frequency crystal oscillators, which can be adjusted to cater for frequency shifts of up to 1 000 c/s. The reason for having different frequencies for the mark and space filters is that during initial laboratory tests it was convenient to feed the demodulation unit with a 4·5/5·0 kc/s two-tone signal. The processes of amplifying and rectifying are similar in both the mark and space branches, except that the mark-branch output corresponding to the signal-on condition is positive and the space-branch output correspond-

ing to the signal-on condition is negative. Each amplifying and rectifying stage consists of a single-valve transformer-coupled amplifier, and a double-diode valve acting as a full-wave rectifier. The rectified output is then applied to a cathode-follower stage, the low output impedance of which is suitable for feeding the 'assessor'. The assessors interpret the rectified outputs relative to appropriate judgment levels, and thus the presence or absence of signal is independently and continuously assessed in each branch. The outputs of the four assessors are then combined; ideally the assessor outputs should be given square-law weighting before combination, but linear combination in a resistive network has been used for simplicity. Since the combined output varies in magnitude as well as in polarity, it is necessary to amplify and limit it before it is sent to line. This is achieved by means of a d.c. output unit which delivers an output of ± 30 volts for a minimum input level of $\pm 0{\cdot}1$ volt. Several monitoring points are available in the demodulation unit, and the performance of each individual branch can be observed. Also, the output of each branch can be switched on or off as required; this facility could be useful in the event of serious interference affecting, for example, the space frequency, leaving the mark frequency clear.

(3.2) Band-Pass Filter

Pulse networks giving a rectangular pulse of current from a capacitor discharged linearly have been used in radar modulators.[7] A very similar network can be employed to provide a linear voltage response to a step voltage input, and design data for such a network are given in Section 9.2. The filters employed for the demodulation unit were designed to have a build-up time of 10 millisec, and the circuit is shown in Fig. 4. L_1C_1, L_2C_2, L_3C_3

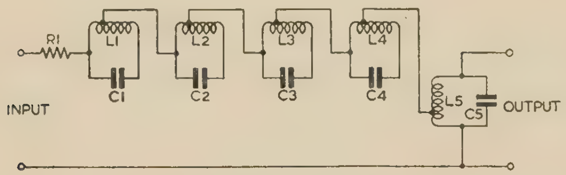


Fig. 4.—5 kc/s band-pass filter.

Tuned circuit	f_0
	c/s
L1C1	5101
L2C2	4901
L3C3	5204
L4C4	4804
L5C5	5000

and L_4C_4 are tuned to produce infinity points at the appropriate frequencies, and L_5C_5 is tuned to the centre frequency of the filter. The response of the filter to a square-wave-modulated signal is shown in Fig. 5, from which it can be seen that the build-up time is approximately 12 millisec instead of the designed value of 10 millisec. Before the initial tests, little time was spent in trying to improve the filter response characteristic, but further experimental work has shown that more accurate alignment can improve the response considerably, and that L_3C_3 and L_4C_4 can probably be dispensed with.

(3.3) Assessor

The circuit diagram of the assessor is shown in Fig. 6. It consists of two rectifiers sharing a common load R_2 and R_3, the charge time-constants C_1R_1 and C_2R_1 being approximately 2 millisec. The discharge time-constant is approximately $C_2(R_2 + R_3)$ when V_1 is conducting, and $C_1(R_2 + R_3)$ when V_2 is conducting. The discharge time-constant for C_1, which is approximately $0{\cdot}4$ sec,

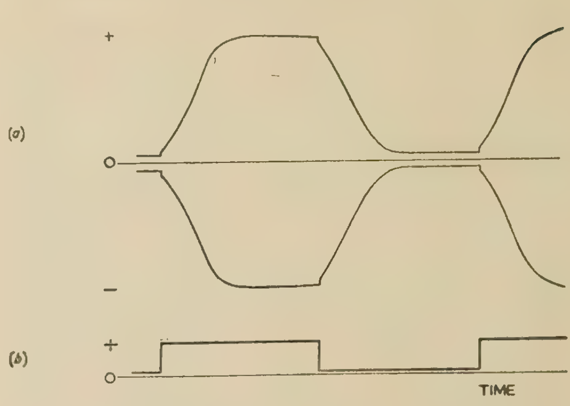


Fig. 5.—Response of filter to step-wave-modulated signal.
(a) Filter output waveform.
(b) Modulating waveform (30-baud reversals).

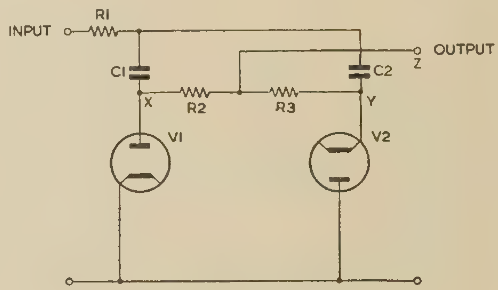


Fig. 6.—Assessor.

is longer than the maximum interval between successive signals of the same kind and shorter than the reciprocal of the fading frequency. Thus in the case of a mark assessor the voltage across C_1 follows the level of interference-plus-desired-signal, and the voltage at X tends to zero during marks and to a negative value (equal to the difference between the level of interference-plus-desired signal and that of interference alone) during spaces. Similarly, the voltage at Y tends to zero during spaces and to a positive value equal to the same difference during marks. The signal output at Z is the average of the signals at X and Y, and is thus positive during marks and negative during spaces, the magnitude being about one-half the difference between the levels of interference and interference-plus-desired-signal. Thus the assessor may be considered to interpret the output of the rectifier in terms of a judgment level roughly half-way between the high level of rectifier output corresponding to the signal-on condition and the low level corresponding to noise and interference only. The judgment level automatically follows the fading of signal and interference, provided that the fading is not too rapid.

(4) TEST RESULTS

(4.1) Laboratory Tests

The new type of demodulation unit has been subjected to comprehensive laboratory tests. Most of the tests were made on the demodulation unit alone, i.e. the filters, rectifiers, assessors and output unit, but a few tests were made to confirm that substantially the same results were obtained when the demodulation unit was fitted with frequency-changing stages and fed from the i.f. amplifier of a typical receiver, as in Fig. 3. The test apparatus and facilities are fully described elsewhere.[8,9] In brief, a telegraph signal generator keys a frequency-shift or two-tone oscillator to produce the desired test signal, which is fed via an

rtificial ionosphere (the fading machine) to the receiver. The eceiver output is regenerated, and errors are identified by comaring the regenerated signal with the transmitted signal, suitably elayed. The number of elements in error and the total number ransmitted are recorded, and the average proportion of elements n error is determined over a time adequate for statistical ability.

Fig. 7 shows the measured performance of the unit on 500 c/s-hift non-fading signals consisting of 100-baud reversals, i.e. lternate marks and spaces, in the presence of white noise. The gnal/noise ratio is defined[4] as the ratio of the effective signal nergy per element, w_e, to the noise power per unit bandwidth, N_0. lso plotted are theoretical curves for an ideal telegraph receiver,[4] or a hypothetical receiver having an exponential error-rate ersus signal/noise-ratio characteristic but the same performance vith fading signals as the ideal,[2] and for an exponential characeristic corresponding to a demodulation factor* of 4 dB with ading signals. It will be observed that the measured perforance closely follows an exponential law. Similar measurements ave been made on a conventional radiotelegraph receiver ppropriately adjusted for the test signal, and these are also

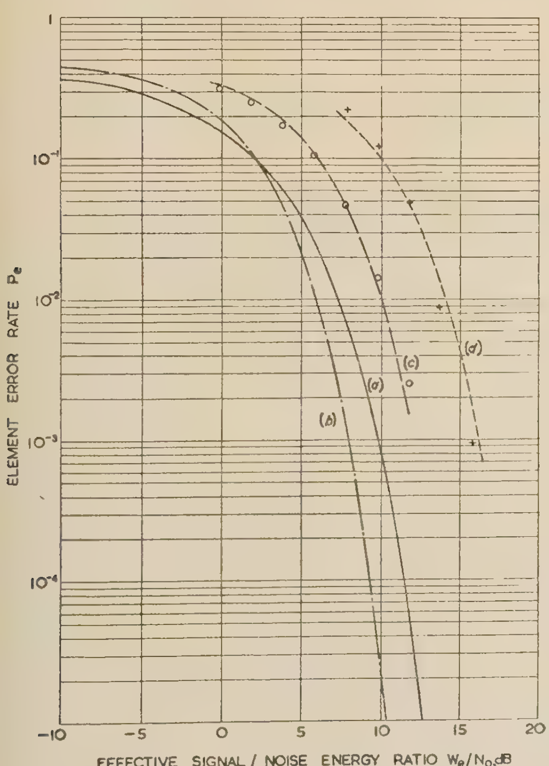


Fig. 7.—Steady-signal performance at 100 bauds.

(a) $P_e = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{w_e}{N_0}\right)^{1/2}$

(b) $P_e = \frac{1}{2}\exp\left(-\frac{w_e}{bN_0}\right)$, where $10 \log_{10} b = 0$ dB

(c) $P_e = \frac{1}{2}\exp\left(-\frac{w_e}{bN_0}\right)$, where $10 \log_{10} b = 4$ dB

(d) $P_e = \frac{1}{2}\exp\left(-\frac{w_e}{bN_0}\right)$, where $10 \log_{10} b = 8$ dB

○ Performance of experimental demodulation unit.
+ Performance of conventional receiver.

plotted in Fig. 7. It will be seen that in this case the performance corresponds to a demodulation factor of about 8 dB on fading signals. Fig. 8 shows how the demodulation factor

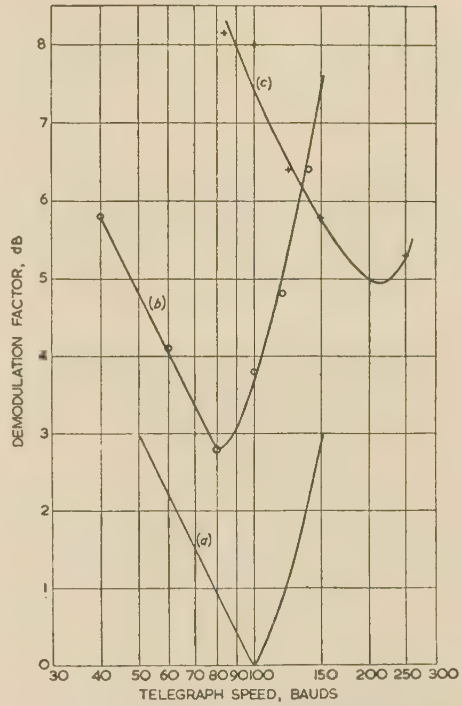


Fig. 8.—Effect of variation of telegraph speed on steady-signal performance.

(a) Performance of ideal receiver designed for 100 bauds.
(b) Performance of experimental demodulation unit.
(c) Performance of conventional receiver.

varies with telegraph speed. It will be seen that the best demodulation factor for the new demodulation unit is about 3 dB, at a speed of 80 bauds, in accordance with the build-up time of the filters. The best result for the conventional receiver, at 220 bauds, is about 5 dB; the designed speed for this receiver, in the condition used, is, in fact, 100 bauds. The significance of the apparent discrepancy is discussed elsewhere.[2] The theoretical curve for a hypothetical receiver designed for ideal performance at 100 bauds can be readily deduced from the proportion of energy in a telegraph element that contributes to the useful output. Such a curve has been plotted in Fig. 8 for comparison. All the curves have the same general shape.

Fig. 9 illustrates the performance of the demodulation unit on a fading signal, the fading frequency being 20 fades/min—somewhat higher than commonly observed in practice. A reversals telegraph signal was used, a speed of 100 bauds being chosen rather than the speed appropriate to the filters, mainly to suit readily available test equipment. The measured points clearly illustrate the improvement in performance between the flat-fading condition resulting from propagation via a single path and the selective-fading condition resulting from propagation via two equal paths having a relative path-time delay of 1 millisec. This latter propagation condition is an optimum for the frequency shift of 500 c/s. Measurements are also shown of the performance at intermediate conditions in which one of the two propagation paths was attenuated relative to the other. The measured fading performance is found to correspond to an average demodulation factor of 3 dB, and the calculated performance corresponding to this demodulation factor has been derived from Fig. 1 and plotted in Fig. 9. Why the demodulation factor differs by 1 dB from that corresponding to the non-fading condition has not been investigated; the discrepancy is possibly due to experimental error. The measured performance of the

* Throughout the paper the term 'demodulation factor' is used for the ratio of the signal/noise ratio of a given receiver to that of an ideal receiver for the same error rate.[4]

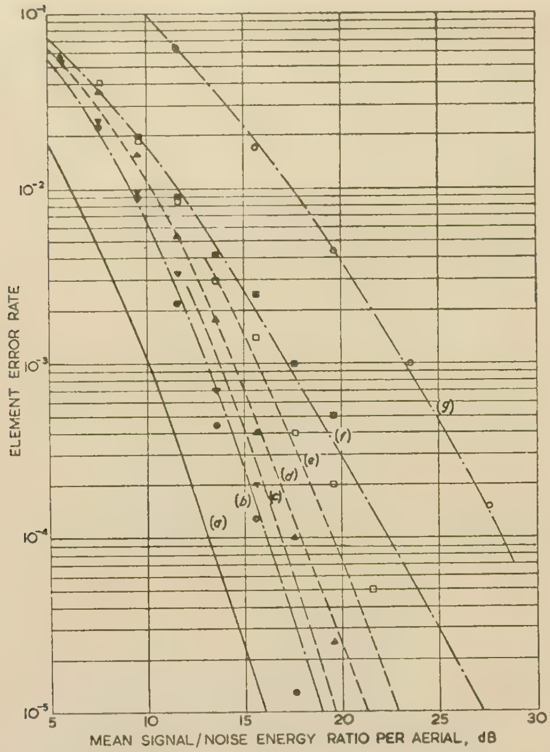


Fig. 9.—Two-path fading-signal performance; 1 millisec relative path-time delay; 100-baud reversals signal; 500 c/s shift.

(*a*) Performance of ideal receiver: equal path attenuations.
(*b*) Performance of a receiver having a demodulation factor of 3 dB; equal path attenuations.
(*c*) As (*b*); relative path attenuation 5 dB.
(*d*) As (*b*); relative path attenuation 10 dB.
(*e*) As (*b*); relative path attenuation 15 dB.
(*f*) As (*b*); single-path propagation.
(*g*) Performance of a receiver having a demodulation factor of 9 dB; single-path propagation.

● Performance of experimental unit; equal path attenuations.
▽ Performance of experimental unit; relative path attenuation 5 dB.
▲ Performance of experimental unit; relative path attenuation 10 dB.
□ Performance of experimental unit; relative path attenuation 15 dB.
■ Performance of experimental unit; single-path propagation.
○ Performance of conventional receiver.

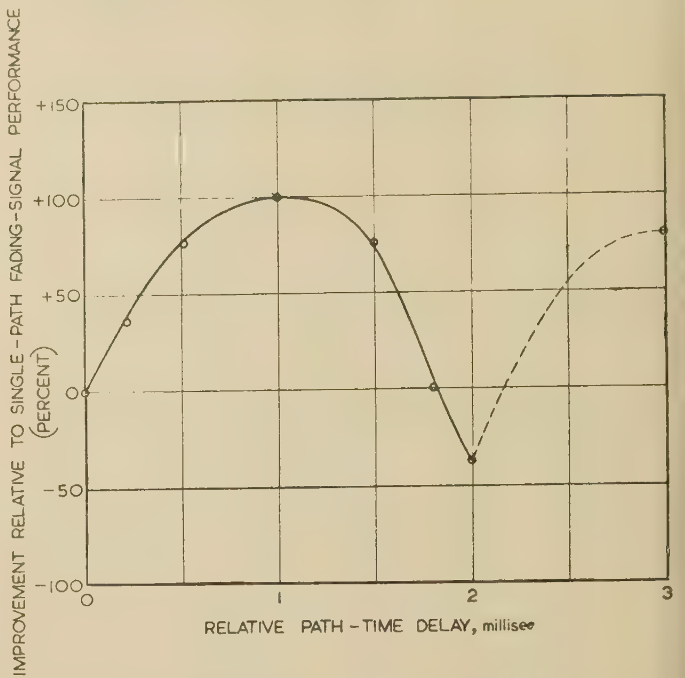


Fig. 10.—Effect of relative path-time delay on two-path fading-signal performance; equal-path attenuations; 100-baud reversals signal; 500 c/s shift.

conventional receiver previously mentioned has also been plotted. It is found to correspond to a demodulation factor of 9 dB under flat-fading conditions; the 1 dB difference from the non-fading performance is to be expected from the fact that diversity switching, rather than combination, is used in this receiver.

The effect has been determined of varying the relative path-time delay in the case of propagation over two equal paths. As might be expected, a family of curves resembling those in Fig. 2 was obtained for delays intermediate between 0 and 1 millisec. The effect of further increase in delay is to worsen the performance until, at a value of 2 millisec, a condition is again reached in which there is complete correlation between the fadings of mark and space signals. Confusion arises from the overlapping of the curves if the method of presentation of Fig. 2 is used. To overcome this, the frequency-diversity improvement, in decibels, for each measurement was expressed as a percentage of the improvement obtained under optimum selective-fading conditions. The improvement for each value of delay was averaged over error rates between 10^{-4} and 10^{-2} and plotted in Fig. 10. It will be seen that the performance at 2 and 3 millisec is considerably worse than that at 0 and 1 millisec respectively. The reason for this is not yet fully understood. A test at 70 bauds, in which the filter build-up time was slightly less than the element length, showed that the performance at 2 millisec was somewhat worse than that at 0 but not to such a marked extent. It therefore appears that it may be best to design the demodulation unit to ignore the periods of distortion that occur at each signal transition under conditions of multi-path propagation. Further investigations are proceeding.

All the tests described so far were made with a reversals telegraph signal. It is to be expected that the new demodulation unit would fail under sustained mark or space signals owing to the appropriate assessor capacitor being discharged and 'forgetting' the conditions that existed in the opposite signal condition. This phenomenon is illustrated in Fig. 11, in which will be seen the considerably worse performance with a 1 : 6 telegraph signal. However, such a signal is unlikely to occur in practice. The results of a test with a random signal, in which the proportion of mark and space signals of different lengths corresponded closely to that resulting from most of the commonly used telegraph codes, will be seen to be similar to those obtained with a reversals signal. The assessor time-constant was chosen arbitrarily, and it may be possible to reduce it without significantly affecting the performance at normal fading rates. If so, it would be desirable, in order to improve the performance of the receiver at more rapid fading rates. It may well be desirable to make the time-constant adjustable in a final design of receiver.

(4.2) Field Tests

For field tests a transmission was required having the following two characteristics:

(*a*) Received signals should exhibit marked selective fading so that the new demodulation unit could demonstrate the frequency-diversity improvement it can give.

(*b*) The stability of mark and space frequencies should be sufficient to permit the narrow filters in the unit being kept accurately in tune.

Measurements show that antipodean signals offer the greatest prospect of continuous selective fading. For example, 80% of

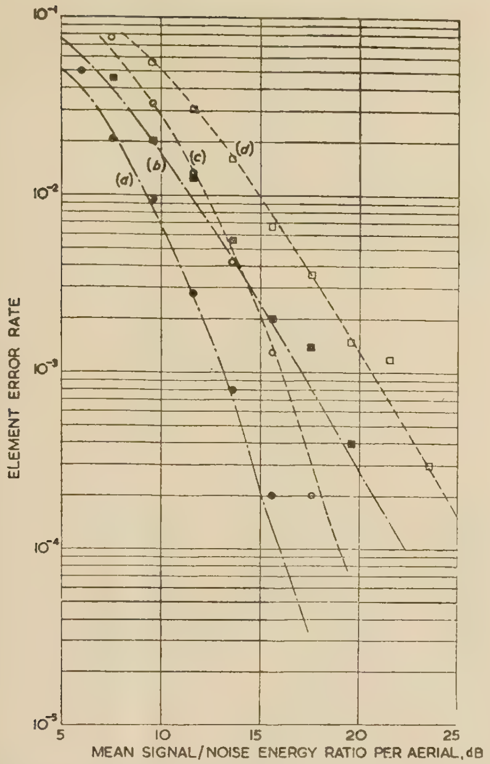


Fig. 11.—Fading-signal performance with various telegraph signal conditions; 100-baud signal, 500 c/s shift.

(*a*) Performance of a receiver having a demodulation factor of 3 dB; optimum two-path propagation.
(*b*) As (*a*) but single-path propagation.
(*c*) As (*a*) but demodulation factor 6 dB.
(*d*) As (*b*) but demodulation factor 6·5 dB.

Performance of experimental unit shown thus:
□ Single-path propagation; 1 : 6 telegraph signal.
■ Single-path propagation; random telegraph signal.
○ Two-path propagation; 1 millisec relative delay; 1 : 6 telegraph signal.
● Two-path propagation; 1 millisec relative delay; random telegraph signal.

facsimile pictures received from Melbourne between October, 1953, and July, 1955, bore signs of distortion due to propagation by several paths of comparable activity, differences in path-time delay being between 0·5 and 2·0 millisec. The frequency-stability requirements of the new method of demodulation have already been discussed; they are considerably more severe than those arising from international regulations. Moreover, it has been found that the frequency-shift stability that is usually realized in current practice is also inadequate. However, the two-channel time-division-multiplex transmission from Melbourne (VIZ 26, 11 660 kc/s) was found to be sufficiently stable in frequency provided that the receiver was re-tuned about every quarter-hour, and the normal aggregate keying speed of 83·6 bauds was suitable for the experimental demodulation unit. The tests were therefore made by means of this transmission. The experimental demodulation unit was taken to Somerton Radio Station and associated with a conventional frequency-shift telegraph receiver of the same type as that used in the laboratory comparisons. The two outputs were fed by line, using a high-speed v.f. telegraph channels, to distributors at the Overseas Telegraph Terminal, Electra House, London, and the A-channel signals from the conventional receiver and from the experimental detection unit were both fed to printers. Specimens of copy were gummed to a sheet every few minutes for analysis.

Comparative tests were made with the signals from the receivers fed straight to line, regeneration taking place only in the distributors at Electra House, and with the signals regenerated before being fed to line at Somerton, to correspond to the conditions of the laboratory tests. It was found that regeneration at the radio station had no significant effect on the performance of either the conventional-receiver channel or the one using the experimental demodulation unit. Evidently, distortion by the line equipment was so small as to have a negligible effect on the total error liability.

Tests were carried out during the periods 18th–19th August and 15th–23rd September, 1955, at a season when the Australian signals were usually very poor for a time at or near mid-day owing to their fading out on the long route over the Pacific and Atlantic Oceans before coming in on the short route over Asia. A typical graph of error rate against time is given in Fig. 12.

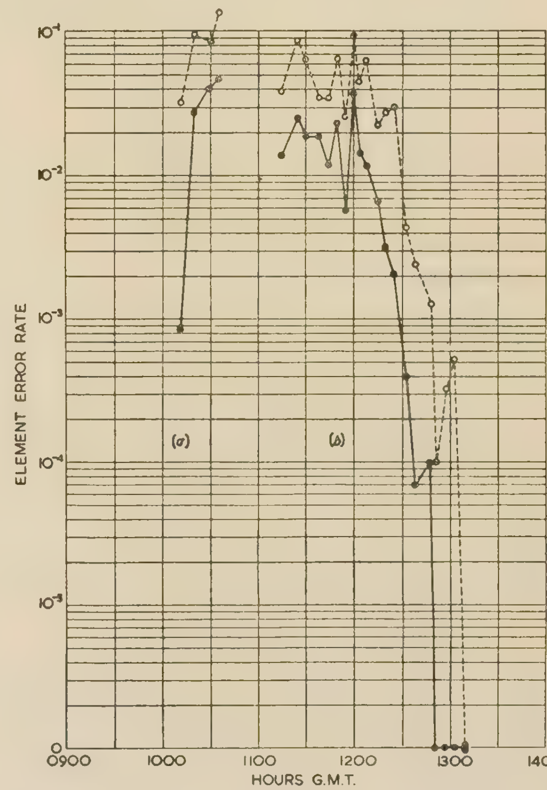


Fig. 12.—Test of experimental receiver on VIZ26 transmission, 20th September, 1955.

(*a*) Fade-out, long route.
(*b*) Fade-in, short route.

---○— Conventional receiver.
—●— Experimental demodulation unit.

For error rates in the range from 1 in 10^3 to 1 in 10^4 the experimental unit generally showed an increase in channel time of the order of from half to one hour during the mid-day period, but there was great variation from day to day. The results of all the comparisons between the experimental and conventional units are given by the points plotted in Fig. 13. Consecutive measurements in which a receiver showed zero errors have been combined with an adjacent measurement having a few errors; for accurate results such low error rates really require long periods of constant signal/noise-plus-interference ratio, which are rarely met in practice. The picture resulting from this arbitrary procedure may be unfair to the experimental unit. The field test results may be summarized by saying that the plotted comparisons lie about half-way between the no-improvement line and the line corresponding to laboratory comparisons under optimum fading conditions. In the important error-rate region around 1 in 10^4

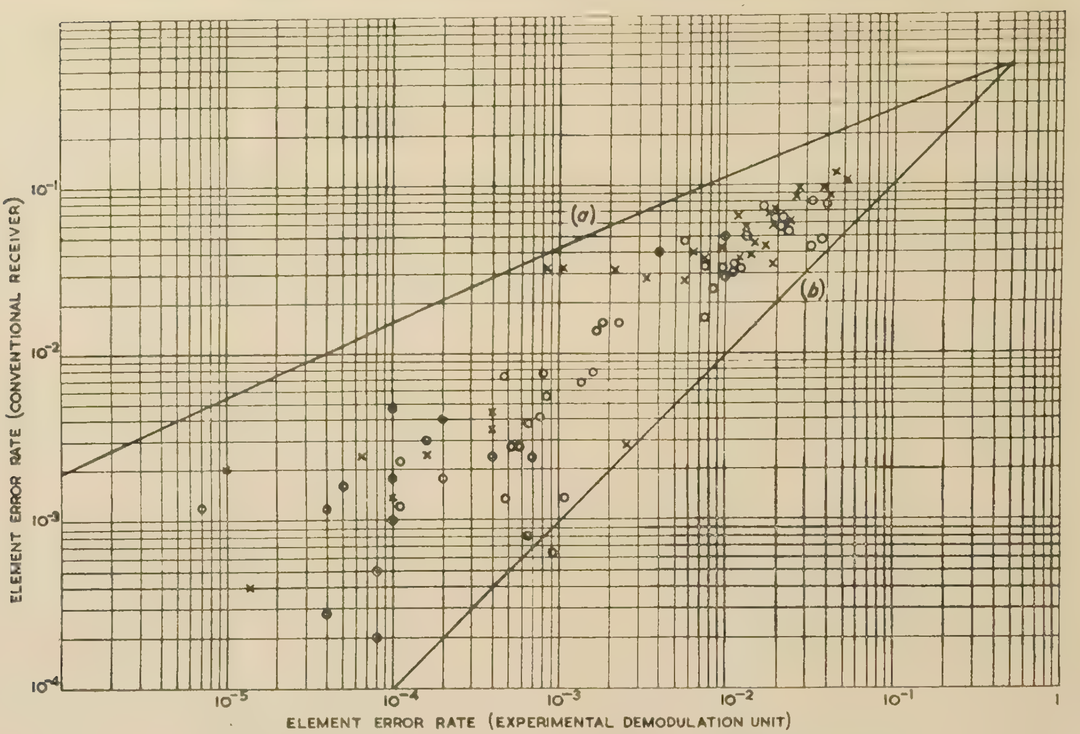


Fig. 13.—Relative performance of conventional and experimental units.

(*a*) Relative performance measured in laboratory in presence of white noise; optimum selective fading; telegraph speed 100 bauds.
(*b*) Curve of equal error rates.
Points are field measurements on VIZ26; 11·66 Mc/s; D.C.C.C., 83·6 bauds; 18th and 19th August and 15th–23rd September, 1955.
○ Experimental unit, output unregenerated.
× Experimental unit, output regenerated.

the experimental unit shows an error liability about ten times better than that of the conventional receiver.

(4.3) Discussion of Test Results

The results of the laboratory tests, in which the signals were disturbed only by white Gaussian noise, show frequency-diversity improvements agreeing well with those calculated from theory. The demodulation factor of 3 dB obtained in the experimental equipment seems satisfactory, considering that the equipment is vulnerable to noise components in quadrature with the signal as well as to in-phase components. By virtue of coherent detection the ideal receiver is not disturbed by the quadrature noise components. Evidently the loss arising from diversity combination on a voltage basis instead of an energy basis was small.

On field trial the experimental unit gave a smaller improvement relative to a conventional receiver than it gave under ideal selective-fading conditions in the laboratory. The results are not easy to interpret in detail. At first sight it might seem that in about 75% of the comparisons the entire improvement could be attributed to the better demodulation factor of the experimental unit, and that frequency diversity made little difference. However, it is known that practical h.f. radio channels are usually more affected by discrete atmospheric crashes and interference than by steady background noise. Some kinds of interference or discrete crashes would, if strong enough, multilate the signals in any receiver, however good, and so tend to produce points on the equal-performance line in comparisons between receivers. It is thought that the improvement due to frequency selectivity may have been diluted in this way. On the other hand, c.w. interference may operate in favour of the new technique, since such interference at a level comparable to the signal and at any frequency within the 1 000 c/s-wide channel-filter pass band of the conventional receiver would cause errors, whereas eqn. (3) suggests that the new unit should only be affected seriously by c.w. interference within about 80 c/s of one or other of the mark and space frequencies. Finally, it is not known how closely the conditions during the field trials approached the ideal of zero correlation between the fadings of mark and space frequencies, nor is it known how well the frequency of fading suited the assessor time-constants.

Further study of all these points is needed.

(5) FURTHER DEVELOPMENTS

There is, doubtless, scope for detailed improvements in the experimental demodulation unit. There is already some laboratory evidence that signals passing through the unit may be unduly disturbed by distorted transitions under multi-path propagation conditions; it may prove desirable to use an integration time somewhat shorter than the nominal duration of an element, even though this must involve some degradation in demodulation factor. Many variations are possible in the arrangement of the assessor. Thus at the expense of complication, which might be justified if rapid fading were to prove common, the judgment level for an element could be determined by reference to later elements as well as earlier ones. On the other hand, some simplification would be possible in a receiver for use exclusively on codes having in every character a number of marks approximately equal to the number of spaces, as in the 4 : 3 error-detecting codes;[10] the judgment level in such a case is very close to the d.c. content of the signal at a detector output, and an assessor might take the form of an a.c. coupling of suitable time-constant.

Useful frequency diversity is only possible if the correlation

of the fadings of mark and space signals is small. Much more information must be obtained about the conditions encountered in practice before desirable values of frequency shift can be assessed; to some extent the correlation may be controllable by choice of working frequency or by the use of special aerial systems capable of selecting particular propagation paths. Very small frequency shifts are, however, clearly undesirable from a frequency-diversity point of view, and the requirements of the new technique may therefore conflict with the need for economy in frequency usage. To the extent that the new technique shows a return in reduced power or higher signalling speed or fewer repetitions, the conflict may be more apparent than real. Relatively large differences between mark and space frequencies combined with frequency economy are possible by interleaving the mark and space frequencies of different channels; a three-channel frequency-division-multiplex system, for example, could have frequencies allocated in the order A-channel mark: B-channel mark: C-channel mark: A-channel space: B-channel space: C-channel space. Transmission in such a system would have to be by the complementary amplitude modulation of mark and space tones and not by frequency modulation.

Methods of coherent detection require investigation to discover whether the modest advantage in signal/noise ratio resulting from immunity to quadrature noise components is worth the complication involved. At the transmitting end, the simplest way of attaining the necessary phase continuity would probably be to use two amplitude-modulated tones derived from continuously running oscillators. The problem is much more difficult at the receiving end. Possibly the simplest solution to the problem of tracking the carrier phase is one suggested by F. J. M. Laver; this is to extract the carrier in each diversity branch by means of a filter having a bandwidth narrow relative to the modulation components but wide relative to the fading frequency. The frequency-stability requirements would be severe.

The new technique depends on the speed of signalling being sufficiently high relative to the speed of fading, and it is not known to what extent this condition is satisfied by the speeds of signalling currently used. It is likely that failures would occur with 'flutter' fading. From this point of view the signalling speed should be as high as the incidence of errors due to multi-path effects allows. Incidentally, high signalling speeds ease the transmitter and receiver frequency-stability requirements and also, in the case of transmissions of the f.m. type with a given frequency shift, make more effective use of the occupied bandwidth.

(6) DISCUSSION AND CONCLUSIONS

The results obtained in the comparison of the experimental unit with a high-grade receiver of the conventional limiter/discriminator type show that the new technique gives a great advantage when the fadings of the mark and space frequencies are uncorrelated and the signal is disturbed only by white Gaussian noise. A smaller, but nevertheless substantial, improvement is obtained under practical working conditions, when signals are more disturbed by interference and atmospheric crashes than by steady noise, and when the relative fadings of mark and space frequencies are not ideal. The advantage of the experimental unit over the conventional receiver stems partly from the utilization of frequency diversity and partly from the reduced effective bandwidth. There is at present insufficient information to assess the relative importance of these contributions under practical working conditions. Further study is required of the characteristics of atmospherics and interference, of the frequency selectivity and speed of fading, and of the effect of varying the equipment design parameters.

An attractive feature of the new technique, as at present applied, is its simplicity. The experimental unit is little more complex than the corresponding parts of the conventional equipment, so that the improvement is obtained cheaply. This is offset to some extent by the need for closer frequency control, though the requirements are not severe if the signalling speed is reasonably high. In view of its cheapness, the early application of the technique to some representative channels would appear worth while, so that the fundamental investigations mentioned in the last paragraph may be supplemented by operational experience.

The experimental work has been confined to the case of long-distance point-to-point radiotelegraphy in the h.f. band at speeds of the order of 100 bauds. The technique may well have wider applications. It would offer particular advantages where site restrictions limit the effectiveness, or even preclude the use, of diversity aerial systems, as in the case of mobile stations operating in the h.f. band. More generally, it is potentially useful in any frequency-modulated data-transmission system with a limited number of signal conditions if the signals are subject to fading that is slow relative to the speed of signalling and is suitably frequency-selective.

The following conclusions may be drawn:

(*a*) In suitable degree, multi-path propagation is to be sought, not avoided.

(*b*) Two-frequency methods of signalling, including frequency modulation, are desirable for the sake of frequency diversity.

(*c*) The ordinary f.m. technique of demodulation, by limiter and discriminator, fails to take advantage of frequency diversity.

(*d*) The improvement conferred by the use of the ordinary f.m. technique of demodulation is of no value if the signals are regenerated before being subjected to further distortion.

(*e*) The ideal receiver would perform a correlation process in each diversity branch, the outputs being weighted according to energy and then combined.

(*f*) Absence of a signal is as significant as presence.

(*g*) Effective frequency diversity can be realized in practice without much complication of the equipment.

(7) ACKNOWLEDGMENTS

Valuable assistance has been given by many members of the staffs of the Post Office Research Station, of the Overseas Telegraph Terminal, and of Somerton Radio Station; in particular, Mr. H. J. Orchard produced the basic design for the filters, Messrs. E. R. Broad and P. F. A. Richards designed the variable-frequency crystal oscillators, and Messrs. W. C. Allen, F. J. Lee and J. D. C. Madder were largely responsible for the arrangement of the field trials. Most of the constructional work and many of the measurements and analyses of test results have been done by Mr. D. L. Lisney. To all these the authors express their gratitude.

Acknowledgment is made to the Engineer-in-Chief of the Post Office and to the Controller of Her Majesty's Stationery Office for permission to publish the paper.

(8) REFERENCES

(1) JOWETT, J. K. S., and EVANS, G. O.: 'A Study of Commercial Time Lost on Transatlantic Radio Circuits due to Disturbed Ionospheric Conditions', *Proceedings I.E.E.*, Paper No. 1708 R, October, 1954 (**102** B, p. 505).

(2) LAW, H. B.: 'The Signal/Noise Performance Rating of Receivers for Long-Distance Synchronous Radiotelegraph Systems using Frequency Modulation' (see page 124).

(3) WOODWARD, P. M., and DAVIES, I. L.: 'Information Theory and Inverse Probability in Telecommunication', *Proceedings I.E.E.*, Paper No. 1239 R, March, 1952 (**99**, Part III, p. 37).

(4) LAW, H. B.: 'The Detectability of Fading Radiotelegraph Signals in Noise' (see page 130).
(5) KAHN, I. R.: 'Ratio Squarer', *Proceedings of the Institute of Radio Engineers*, November, 1954, 42, p. 1704.
(6) DOELZ, M. L.: 'Predicted-Wave Radioteleprinter', *Electronics*, December, 1954, **42**, p. 166.
(7) BOWLER, G. E., MCILROY, M. S., and COATE, G. T.: 'Principles of Radar', Second Edition (McGraw-Hill, 1946), Chapter 6, Art. 4.
(8) HILTON, C. G., LAW, H. B., LEE, F. J., and LEVETT, F. A. W.: 'Laboratory Test Equipment for Synchronous Regenerative Radiotelegraph Systems' (see page 141).
(9) LAW, H. B., LEE, F. J., LOOSER, R. C., and LEVETT, F. A. W.: 'An Improved Fading Machine' (see page 117).
(10) HAYTON, T., HUGHES, C. J., and SAUNDERS, R. L.: 'Telegraph Codes and Code Convertors', *Proceedings I.E.E.*, Paper No. 1585 R, November, 1953 (**101**, Part III, p. 137).

(9) APPENDICES

(9.1) Performance of an Ideal Receiver Fed with Rayleigh-Fading Signals and Noise under Two-Path Propagation Conditions

Suppose that a two-path propagation condition exists whereby an aerial is receiving an amount of power p_a via one propagation path and an amount p_b via the other, and that these signal components fade independently according to the Rayleigh distribution, which may be written in the form

$$\mathrm{P}(p) = 1 - \exp(-p/P) \quad . \quad . \quad . \quad . \quad . \quad (6)$$

where $\mathrm{P}(p)$ is the proportion of time for which the signal power is less than p, and P is the mean signal power. Suppose, further, that the signal is of the frequency-shift-telegraphy type, with a frequency shift Δf, and that the difference between the path-time delays of the two paths is τ. The phase relations between the shorter-path and the longer-path components will differ by

$$2\pi\tau\Delta f = \phi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

for the mark and space frequencies. Thus we may write for the mark-frequency power, p_m, when mark is present at a given instant

$$p_m = p_a + p_b - 2(p_a p_b)^{1/2}\cos\left(\theta - \frac{\phi}{2}\right) \quad . \quad . \quad (8)$$

and similarly for the space signal power, p_s, if space is present at the given instant

$$p_s = p_a + p_b - 2(p_a p_b)^{1/2}\cos\left(\theta + \frac{\phi}{2}\right) \quad . \quad . \quad (9)$$

where θ varies with time. All values of θ are equally likely, since p_a and p_b fade independently. Thus, apart from the special cases $\phi = 0, 2\pi$, etc., the resultant mark and space powers vary differently with time and there is the possibility of a frequency-diversity advantage.

In the case of steady signals it has been shown[4] that the error liability of an ideal diversity system in the presence of white Gaussian noise can be expressed in terms of the effective signal energy per element, w_e, defined as half the sum of the mark-signal-element and space-signal-element energies received when mark and space respectively are sent. Thus for the case of equal *a priori* probabilities of mark and space we have

$$\mathrm{P}_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\,(w_e/N_0)^{1/2} \quad . \quad . \quad . \quad . \quad (10)$$

where P_e is the probability of a binary-code signal element being received wrongly and N_0 is the noise power per unit bandwidth, assumed equal in all diversity branches. If the fading is very slow compared with the speed of signalling the signal power may be regarded as constant in any one signal element, so that eqn. (10) holds, and we may take account of the fading as a variation from element to element in the amount of energy received. To determine the effective signal energy in the single-aerial frequency-diversity case we must take half the sum of the mark and space energies; thus, adding eqns. (8) and (9) and rewriting in terms of energy,

$$\begin{aligned} w_e &= \tfrac{1}{2}(w_m + w_s) \\ &= w_a + w_b - (w_a w_b)^{1/2}\left[\cos\left(\theta - \frac{\phi}{2}\right) + \cos\left(\theta + \frac{\phi}{2}\right)\right] \\ &= w_a + w_b - 2(w_a w_b)^{1/2}\cos\theta\cos\frac{\phi}{2} \quad . \quad . \quad . \quad . \quad . \quad (11) \end{aligned}$$

Two special cases may be noted. If $\phi = 0$, we have

$$w_e = w_a + w_b - 2(w_a w_b)^{1/2}\cos\theta \quad . \quad . \quad . \quad (12)$$

which, since θ may take any value, is simply a matter of adding two components in random phase. Since the components individually conform to the Rayleigh distribution, the sum will also conform to the same distribution; the fading is not frequency-selective and there is no possibility of a frequency-diversity advantage.

If $\phi = \pi$, we have

$$w_e = w_a + w_b \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

In this case the random phase θ has disappeared from the expression and the effective signal energy is always equal to the sum of the energies of the two components. The distribution for the sum is more favourable than the Rayleigh distribution, and, if the mean values of w_a and w_b are equal, i.e. if the two propagation paths are equally active, there is zero correlation between the fadings of the mark and space signals and the full dual-diversity advantage is possible.

By determining the probability density of effective signal energy and applying eqn. (10) the probability of error in the general two-path case could be determined. The general solution has not been obtained, but an idea of the behaviour of an ideal receiver under two-path-propagation conditions has been obtained by treating some special cases.

(9.1.1) Ideal Path-Time-Delay Difference—Single Aerial.

The ideal path-time-delay difference gives $\phi = \pi$, so that eqn. (13) holds. The energies w_a and w_b conform to the Rayleigh distribution, and for their probability densities we have, from eqn. (6),

$$\mathrm{p}(w_a) = \frac{1}{W_a}\exp(-w_a/W_a) \quad . \quad . \quad . \quad . \quad (14)$$

$$\mathrm{p}(w_b) = \frac{1}{W_b}\exp(-w_b/W_b) \quad . \quad . \quad . \quad . \quad (15)$$

For the probability density of the effective signal energy we have

$$\begin{aligned} \mathrm{p}(w_e) &= \int_0^{w_e}\frac{1}{W_a}\exp(-w_a/W_a)\frac{1}{W_b}\exp\left(-\frac{w_e - w_a}{W_b}\right)dw_a \\ &= \frac{1}{W_a - W_b}\left[\exp(-w_e/W_a) - \exp(-w_e/W_b)\right] \quad . \quad (16) \end{aligned}$$

The probability of error, P, for given values of W_a and W_b can now be obtained by the integration

$$P = \int_0^\infty p(w_e)P_e dw_e \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

Numerical integrations have been performed for values of W_a/W_b of 5, 10 and 15 dB, with the results shown in the single-aerial curves in Fig. 1. The signal/noise energy-ratio scale is in terms of the mean energy per signal element received by the aerial, W_1, given by

$$W_1 = W_a + W_b \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

The case $W_a = W_b = \frac{1}{2}W_1$ gives, in place of eqn. (16),

$$p(w_e) = \int_0^{w_e} (2/W_1)^2 \exp(-2w_e/W_1)dw_a$$

$$= w_e(2/W_1)^2 \exp(-2w_e/W_1) \quad . \quad . \quad . \quad (19)$$

which, apart from minor differences in notation, differs from eqn. (18) of Reference 4 only in respect of the factor $\frac{1}{2}$ associated with W_1. The results given in that paper can therefore be applied, with a 3 dB correction, to the present case. The reason for the 3 dB change is that in the previous paper the results were expressed in terms of the signal energy per diversity branch, whereas here they are in terms of the energy per aerial, i.e. per two branches (mark and space). The case $W_a = W_1$, $W_b = 0$ corresponds precisely to the no-diversity case of the previous paper. The results for $W_a = W_b$ and $W_b = 0$ are also plotted in Fig. 1.

(9.1.2) **Ideal Path-Time-Delay Difference—Two Aerials.**

If we have two aerials, aerial 1 and aerial 2, both of which receive the same mean signal energy from the two paths, but which are far enough apart for the signal-strength variations in them to be uncorrelated, the probability density, $p(w_1)$, of the signal energy received by aerial 1 is, by eqn. (16),

$$p(w_1) = \frac{1}{W_a - W_b}[\exp(-w_1/W_a) - \exp(-w_1/W_b)] \quad . \quad (20)$$

Similarly

$$p(w_2) = \frac{1}{W_a - W_b}[\exp(-w_2/W_a) - \exp(-w_2/W_b)] \quad . \quad (21)$$

The total effective energy is now given by

$$w_e = w_1 + w_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

and for the probability density of w_e we have

$$p(w_e) = (W_a - W_b)^{-2}\int_0^{w_e}\left[\exp\left(-\frac{w_1}{W_a}\right) - \exp\left(-\frac{w_1}{W_b}\right)\right] \times$$

$$\left[\exp\left(-\frac{w_e - w_1}{W_a}\right) - \exp\left(-\frac{w_e - w_1}{W_b}\right)\right]dw_1$$

$$= (W_a - W_b)^{-2}\left\{w_e \exp\left(-\frac{w_e}{W_a}\right) + w_e \exp\left(-\frac{w_e}{W_b}\right)\right.$$

$$\left.-\frac{2W_aW_b}{W_a - W_b}\left[\exp\left(-\frac{w_e}{W_a}\right) - \exp\left(-\frac{w_e}{W_b}\right)\right]\right\} \quad . \quad (23)$$

Eqn. (17) may now be applied. Numerical integrations have been performed for $W_a/W_b = 5$, 10 and 15 dB with the results shown, in terms of the mean signal energy received by one aerial, in the two-aerial curves of Fig. 1.

The case $W_a = W_b = \frac{1}{2}W_1$ gives

$$p(w_e) = \tfrac{1}{6}w_e^3(2/W_1)^4 \exp(-2w_e/W_1) \quad . \quad . \quad . \quad (24)$$

which result, like the single-aerial one obtained already, differs from the corresponding one in the previous paper only by a factor of 3 dB. The case $W_a = W_1$, $W_b = 0$ corresponds precisely to the dual-diversity case of the previous paper.

(9.1.3) **Propagation Paths Equally Active.**

The general expression for the effective signal energy in one aerial, eqn. (11), may be rewritten in the form

$$w_e = (w_a + w_b)\left(1 - \cos\frac{\phi}{2}\right) + [w_a + w_b - 2(w_aw_b)^{1/2}\cos\theta]\cos\frac{\phi}{2}$$

$$= w_d + w_v \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

where w_d is a diversity component given by

$$w_d = (w_a + w_b)\left(1 - \cos\frac{\phi}{2}\right) \quad . \quad . \quad . \quad (26)$$

and w_v is a Rayleigh-fading component given by

$$w_v = [w_a + w_b - 2(w_aw_b)^{1/2}\cos\theta]\cos\frac{\phi}{2} \quad . \quad . \quad (27)$$

Both these components are essentially positive, and the distribution of the diversity component is more favourable than the Rayleigh distribution in the sense that low levels are less probable. It follows that the overall error liability will not be worse than (*a*) that corresponding to the complete signal fading Rayleigh-fashion, or (*b*) that due to the diversity component alone, whichever is the less. We are thus able to determine upper limits to the error liability.

Taking the case of two equally active propagation paths, so that $W_a = W_b = \frac{1}{2}W_1$, the probability density given in eqn. (19) applies, and when $\phi = \pi$ the error liability curve already obtained results. When $\phi \neq \pi$ the energy of the diversity component w_d is abated in proportion to $\left(1 - \cos\frac{\phi}{2}\right)$ and a corresponding adjustment to the error-liability curve is required. When $\phi = 0$ we have the no-diversity case again.

Similar considerations apply in the two-aerial case. The error-liability curve for $\phi = \pi$ has already been calculated, and the abating factor when $\phi \neq \pi$ is the same as in the single-aerial case.

Curves for $\phi = 0$, 30, 60, 90 . . . 180° are given in Fig. 2 for the cases of one and two aerials. As indicated above, these curves will represent upper limits to the error liability; it is to be expected that the complete curve for a given value of ϕ will lie along the corresponding given curve when the error liability is low, and that as the probability of error increases it will eventually merge smoothly with the $\phi = 0$ curve.

(9.2) Design Data for Band-Pass Filter

To obtain a linear build-up response to a step-wave the required transfer function (output voltage)/(input voltage), normalized for a build-up time of 2π is given by the formula $[1 - \exp(-2\pi\nu)](2\pi\nu)^{-1}$, where ν is the complex frequency. It can be shown that

$$\frac{1 - \exp(-2\pi\nu)}{2\pi\nu} = \frac{1}{\pi\nu(1 + \coth\pi\nu)}$$

also

$$\coth\pi\nu = \frac{1}{\pi\nu} + \sum_{n=1}^{\infty}\frac{2\pi\nu}{(\nu^2 + n^2)\pi^2}$$

where n is the order of approximation.

Thus, by arranging the input signal, which is assumed to be obtained from a constant-voltage source, to be fed into a series network, the impedance of which is $1 + \frac{1}{\pi\nu} + \sum_{n=1}^{\infty} \frac{2\pi\nu}{(\nu^2 + n^2)\pi^2}$ and tapping off the output from the component $1/\pi\nu$, the desired response can be realized. In practice, the first component of the impedance corresponds to a resistor of 1 ohm, the second to a capacitor of π farads and the third to an infinite series of parallel-tuned circuits having capacitors of $\pi/2$ farads and inductors of $2/\pi$, $2/4\pi$, $2/9\pi$, . . ., etc., henrys. It will be observed that the effect of the successive tuned circuits becomes progressively less. Thus any desired order of approximation is obtained by using an appropriate number of circuits.

The above arrangement is, of course, a low-pass filter. To obtain a band-pass filter having the equivalent envelope response the appropriate low-pass to band-pass transformation must be applied. This involves replacing the capacitor by a tuned circuit and replacing each of the original tuned circuits by a pair of tuned circuits, the arrangement being that of Fig. 4. The values of L_5C_5 corresponding to a band-pass central angular frequency of unity are given by $1/\pi k$ and πk respectively, where k is half the product of the central frequency and the build-up time. The values of L_1C_1, L_2C_2, etc., are given for successive orders of approximation by

$$L_l = \frac{2}{\pi(\omega_l + \omega_h)k\omega_l} \qquad L_h = \frac{2}{\pi(\omega_l + \omega_h)k\omega_h}$$

$$C_l = \frac{\pi k(\omega_l + \omega_h)}{2\omega} \qquad C_h = \frac{\pi k(\omega_l + \omega_h)}{2\omega_h}$$

where

$$\omega_l = [(n/2k)^2 + 1]^{1/2} - n/2k \qquad \omega_h = [(n/2k)^2 + 1]^{1/2} + n/2k$$

For central frequencies, ω_0, other than unity, and where, as would normally hold, it is desired to employ a generator having an impedance, R_0, other than 1 ohm, it is necessary to multiply all inductor values by the factor R_0/ω_0 and all capacitor values by $1/\omega_0 R_0$.

[The discussion on the above paper will be found on page 147.]

621.376.3

Paper No. 2152 R
Aug. 1956

AN INVESTIGATION OF THE SPECTRA OF BINARY FREQUENCY-MODULATED SIGNALS WITH VARIOUS BUILD-UP WAVEFORMS

By J. W. ALLNATT, B.Sc.(Eng.), Associate Member, and E. D. J. JONES.

(*The paper was first received* 7*th March, and in revised form* 10*th May*, 1956. *It was published in August*, 1956, *and was read before the* RADIO AND TELECOMMUNICATION SECTION 14*th November*, 1956.)

SUMMARY

It is found that a straight-line waveform is the best that can be employed for the transition between mark and space frequencies in frequency-shift radiotelegraphy, since this waveform involves the minimum occupied bandwidth for a given total build-up time. Data on the spectra of frequency-shift signals using straight-line build-up, i.e. a trapezoidal waveform, are presented for modulation indices between 1 and 6 and build-up times between 0 and 0·31 of a telegraph element. Although the work was initiated for radiotelegraphy, the results are, of course, applicable to frequency-modulated signals generally.

(1) INTRODUCTION

Radiocommunication in the 4–30 Mc/s band is often impeded by interference from unwanted transmissions, and there is thus a clear need to restrict transmission bandwidths.[1] An obvious method of bandwidth limitation is by filtering, as exemplified in independent-side-band telephony, where the signal is usually filtered at frequencies of about 100 kc/s and then passed through linear frequency-changers and amplifiers. However, in single-channel-telegraphy transmitters, class C amplifiers are often used on the score of simplicity and efficiency. Filtering of the modulated signal can then be effective only if it follows the amplifiers at a point where the signal has such high power and frequency as to render filter design impracticable. Frequency modulation offers an alternative method of bandwidth limitation by control of the modulating waveform, by filtering or otherwise.

In general, slowing down the build-up time of the modulating signal reduces the bandwidth, but it also makes interpretation of the received signal more difficult, so that a compromise must be sought.[2] The waveform of the build-up has a secondary effect on the bandwidth, and the best waveform may evidently be defined as that which produces the minimum bandwidth for a given total build-up time.

(2) GENERAL

Little information is available upon the spectra of signals frequency-modulated with any but the most simple waveforms. Cawthra and Thomson have given a very complete theoretical treatment[3] for rectangular-wave modulation with half-sine-wave build-up, which, of course, includes plain rectangular-wave modulation as a limiting case. A contribution from Japan to the VIIth Plenary Assembly of the C.C.I.R. (1953) presented experimental data supported by an approximate theoretical analysis for trapezoidal-wave modulation. From these works it was evident that a theoretical treatment of the effect of altering the build-up waveform would be very complex, and it was therefore decided to make a purely experimental investigation, comparison with existing data being made whenever possible to check the validity of the results.

The following salient characteristics relating to half-sine-wave build-up are quoted from the work of Cawthra and Thomson.[3]

Mr. Allnatt and Mr. Jones are at the Post Office Research Station.

The spectrum of a sinusoidal carrier frequency-modulated with a uniform rectangular wave having zero build-up time may be stated thus:

$$A_r = \frac{2m}{\pi(r^2 - m^2)} \sin (r - m)\frac{\pi}{2}$$

where A_r = Amplitude of the rth component.
m = Modulation index, as defined in C.C.I.R. Recommendation No. 87.[4]
r = Component number of spectrum, the number of the component at the mid- (or carrier-) frequency being taken as zero.

It should be noted that m as here defined has one-half the value used by Cawthra and Thomson.

It will be seen that the spectrum is bounded, in general, by an envelope given by $2m/\pi(r^2 - m^2)$. If m is an odd integer, even components coincide with the envelope whilst odd components (other than when $r = m$) are zero. If m is an even integer, odd components coincide with the envelope whilst even components (other than when $r = m$) are zero. If $2m$ is an odd integer, all components become equal to $1/\sqrt{2}$ of the envelope value. In the general case, for intermediate values of m, all components lie below the envelope, varying cyclically, each cycle comprising two components.

When the build-up time is finite, the same cyclic characteristic is maintained in the spectrum but the expressions both for the amplitude of individual components and for the envelope become very complex. Curves, normalized with respect to m, of the envelope for various build-up times are presented in Fig. 1(*a*). It should be noted that the value of build-up ratio, s, used here and elsewhere is twice that used in Reference 3. Also, the definition of build-up time used is the total time for transition between mark and space frequencies, in contrast with the 10%–90% build-up time defined in Reference 4. Multipath propagation can cause distortion of F1-signal elements during the time that different frequencies are simultaneously present at the receiver input. In this respect, the 0–10% and 90–100% portions of the transition are as important as the central 10–90% portion; the 0–100% definition of build-up time therefore seems more appropriate.

The waveform in Fig. 1(*a*) corresponds to a telegraph signal, usually known as a reversals signal, consisting of alternate marks and spaces, each having a duration of one telegraph element, transmitted at a telegraph speed of twice the modulation repetition frequency. If the durations of the mark and space components of the modulating wave are unequal the envelope remains unchanged but the cyclic factor in the individual components has a length of $\frac{\text{(mark time)} + \text{(space time)}}{\text{(mark or space time, whichever is the shorter)}}$ components. In normal telegraph traffic the duration of marks and spaces will often be greater than one element, but it can be shown that the spectrum amplitude never exceeds that for a reversals signal. For this reason all the tests are made with reversals signals.

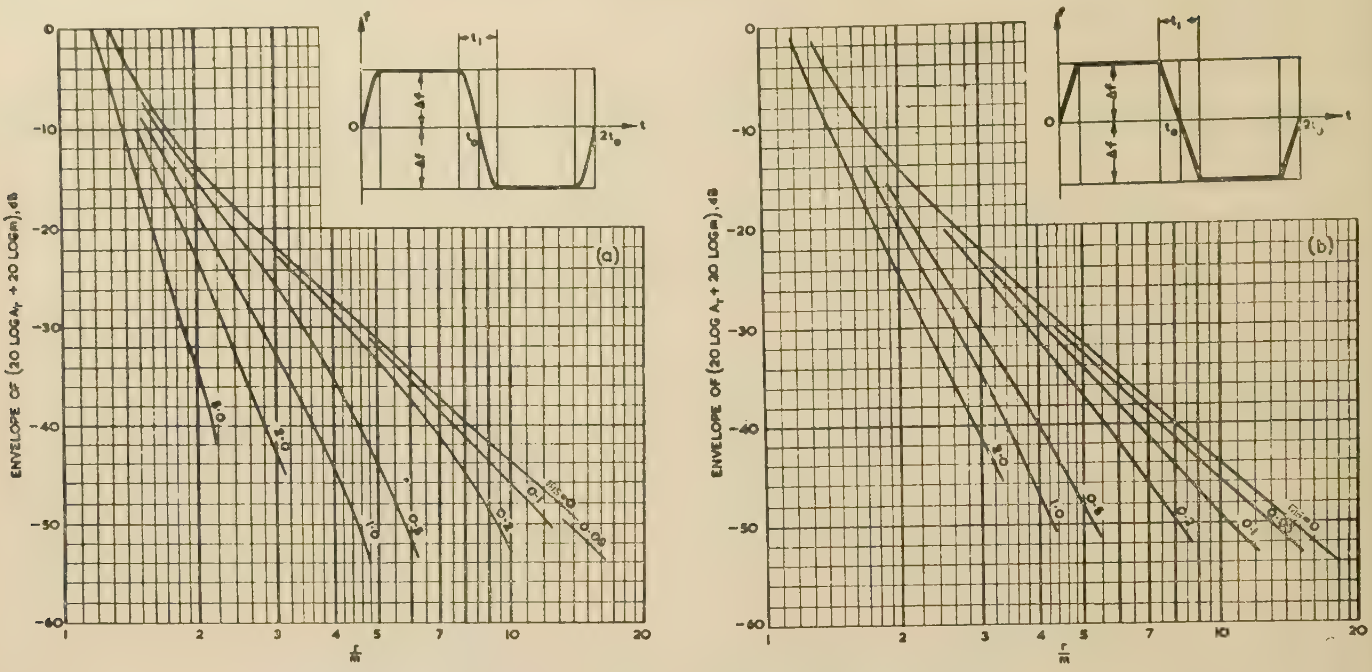


Fig. 1.—Spectrum of frequency-modulated sinusoidal carrier wave.

(*a*) Modulation by a rectangular waveform with half-sine-wave build-up (calculated). These curves may also be applied to signals having mark/space ratios other than unity. In this case t_0 should be taken as the average of mark and space durations.

(*b*) Modulation by a trapezoidal waveform (empirical).

t_0 = Duration of telegraph element.
$f_r = 1/2t_0$ = Modulation repetition frequency.
$2\Delta f$ = Overall frequency shift.

$m = \Delta f/f_r$ = Modulation index.
t_1 = Overall build-up time.
$s = t_1/t_0$ = Build-up ratio.

A_r = Spectrum envelope amplitude at frequency corresponding to rth component.

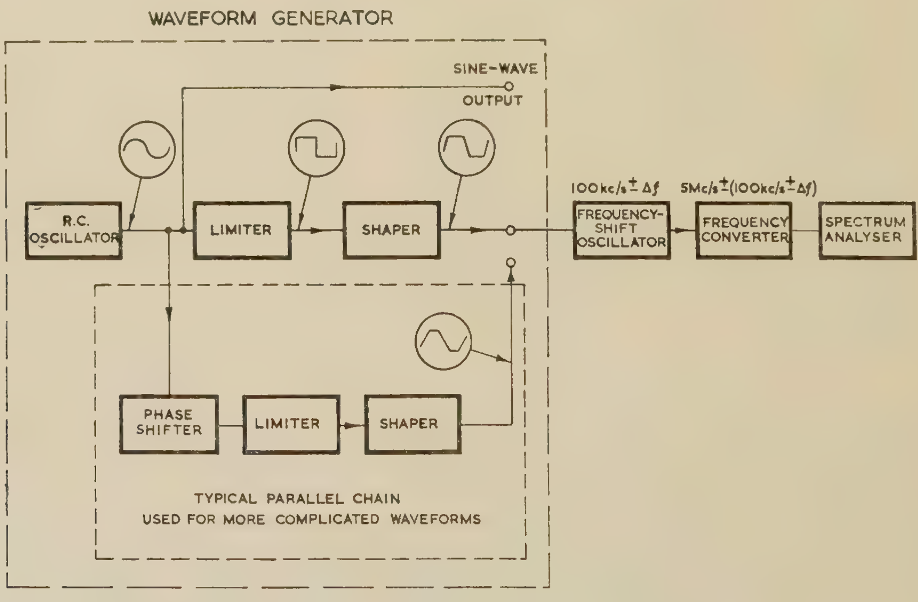


Fig. 2.—Equipment for measurement of spectra of f.m. signals.

(3) EQUIPMENT USED FOR TESTS

The block schematic in Fig. 2 gives the main essentials of the test equipment and method of connection. A low-frequency waveform generator capable of providing a variety of waveforms was used to frequency-modulate a 100 kc/s oscillator, the spectrum of the resulting signal being continuously scanned by a spectrum analyser and displayed on a cathode-ray tube.

The low-frequency oscillator in the waveform generator was a Wien-bridge circuit giving a sinusoidal output of about 50 volts peak-to-peak at 100 c/s, i.e. at a telegraph speed of 200 bauds. This output was available as a modulating signal when required and was also fed to a 2-stage limiter adjusted to give square waves of about 180 volts peak-to-peak with equal mark-to-space ratio. By applying the square wave to a shaper comprising an *RC* integrating circuit of appropriate time-constant, the voltage across the capacitor being limited to ± 10

volts by a double-diode limiter circuit, a trapezoidal wave was obtained which was fed to the output via a cathode-follower. Sine, square, and trapezoidal waveforms of the same frequency were thus available, adjustable in amplitude by potentiometers. More complicated waveforms could also be obtained by combining two or three trapezoidal waves in the appropriate relative phase. A typical arrangement for superposition of one additional trapezoidal wave is shown in the lower part of Fig. 2.

The frequency-shift oscillator, which consisted of a reactance valve and an oscillator valve, delivered an output with an available linear frequency shift of 1 kc/s centred on 100 kc/s so that a modulation index of 10 could be obtained without distortion. A carrier of about 5 Mc/s was amplitude-modulated with the frequency-modulated 100 kc/s signal, since the working range of the spectrum analyser was 3–30 Mc/s. The spectrum analyser was the prototype of an instrument described elsewhere.[5] It displayed the carrier and its associated sidebands on a 6 in-diameter tube with a long after-glow, and sideband levels down to −60 dB relative to carrier level could be read directly from the screen. As used in the tests, the analyser had a bandwidth of 6 c/s and was adjusted to scan a range of about 3 kc/s once in ten seconds.

(4) METHOD USED FOR MEASURING VARIOUS SPECTRA

The following method of adjusting the depth of modulation of the signal was adopted for each test. With the aid of Bessel function curves the spectrum for the desired modulation index with sine-wave modulation was deduced. Using sine-wave modulation, the depth of modulation was adjusted to give this spectrum. At the sweep speed used, there was a certain amount of ringing of the 6 c/s filter in the spectrum analyser. The effect of this was not serious, consisting mainly of a slight reduction in the amplitude of all components, and was tolerated to reduce the number of adjustments to the spectrum-analyser frequency needed to cover a given band. It was found possible, in all cases, to adjust the depth of modulation so that the measured amplitudes were within ±1 dB of the calculated amplitudes for all components down to −60 dB relative to carrier level.

The waveform under test was then substituted for the sine wave and the resulting spectrum was recorded; by arranging that the peak-to-peak voltage of the waveform under test was the same as that of the calibrating sine wave, the required modulation index was obtained. The measured spectrum envelopes with square-wave modulation agreed to within ±2 dB with those obtained mathematically, confirming that the methods of adjustment and measurement were reasonably accurate and that confidence could be placed in the results obtained with the more complicated waveforms, which were, in general, intermediate between a square wave and a sine wave.

As shown in Section 2, integral values of modulation index should lead to the disappearance of alternate spectrum components; their presence in practice and their lack of symmetry about the carrier were probably due to slight errors in the adjustment of the depth of modulation and assymmetry in the modulating waveforms. In general, they were at least 10 dB below the spectrum envelope; if due to incorrect setting of the depth of modulation, they would correspond to errors of less than ±0·1 in the modulation index. Their cause was not investigated, in view of the fact that the envelope amplitudes were not seriously affected.

(5) TESTS TO DETERMINE THE INFLUENCE OF BUILD-UP WAVEFORM ON SPECTRUM

The simplest waveform for a modulating signal with a given build-up time is, of course, a straight line, as illustrated in

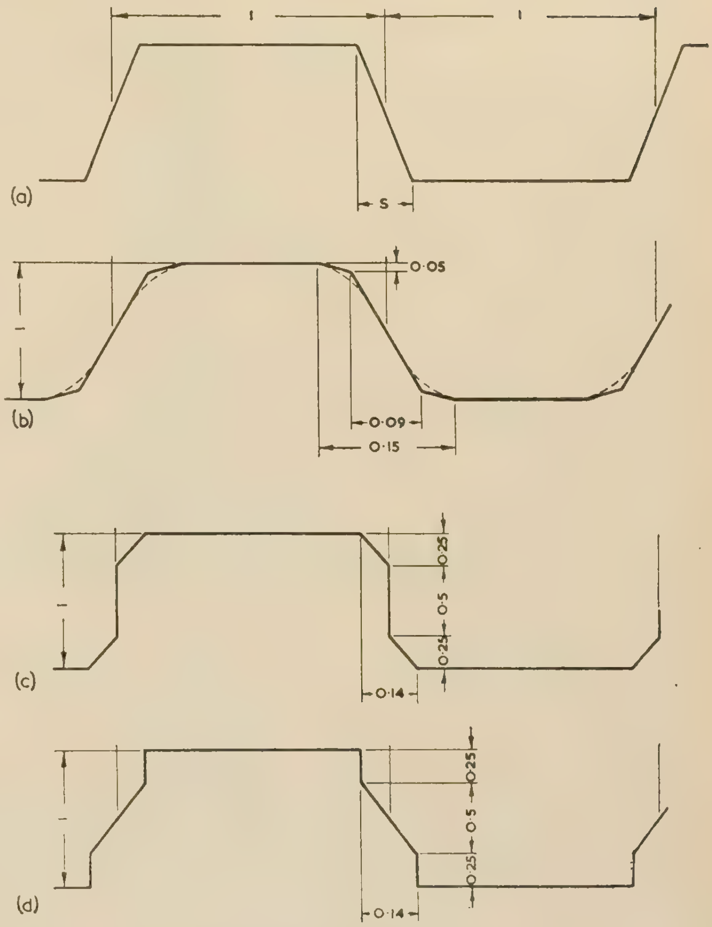


Fig. 3.—Details of waveforms used in tests.

(*a*) Straight-line build-up (trapezoidal waveform).
(*b*) Approximate half-sine-wave build-up.
True half-sine-wave build-up shown thus – – – –
(*c*) Modified rectangular waveform A.
(*d*) Modified rectangular waveform B.

Fig. 3(*a*). Because of the ease of generating such a trapezoidal waveform, it was used for the first series of tests. Fig. 4 gives typical spectra obtained with build-up ratios, s, of 0 to 0·31 and a modulation index, m, of 2; the spectrum due to sinusoidal modulation is included for comparison. Tests were also made with modulation indices of 1, 3 and 6, and the results of all tests are tabulated in Table 1. As an additional check, the sum of the individual component powers was calculated in each test and compared with the unmodulated carrier power. The two powers should, of course, be equal, and in fact were found to be within −0·1 and +0·8 dB of equality. Further tests were carried out with a trapezoidal wave having $s = 0·11$ using modulation indices of 1, 1·2, 1·5, 1·8 and 2. The results, in Fig. 5, show clearly the contribution of the different components to the spectra, and how alternate components disappear for integral values of modulation index.

Having determined the spectral distribution for simple trapezoidal modulation, the next step was to consider other likely waveforms. Amplitude modulation by a trapezoidal waveform, as is well known, causes a wide frequency spectrum arising from the second-order discontinuities at the corners of the waveform, and it seemed possible that these discontinuities might contribute significantly to the bandwidth of the corresponding frequency-modulated signal. The half-sine-wave build-up discussed in Section 2 avoids these discontinuities,

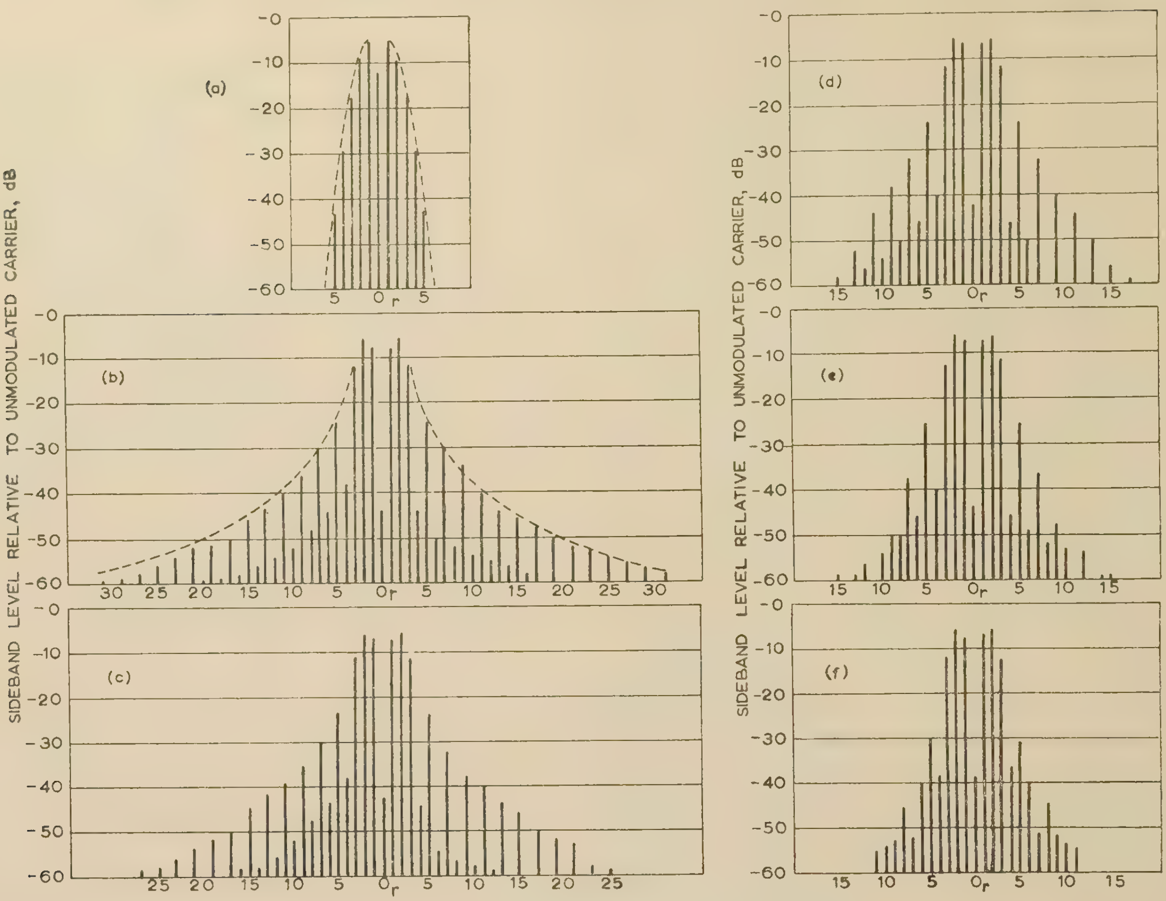


Fig. 4.—Typical measured spectra of frequency-modulated signals, $m = 2$.

Calculated spectrum envelopes shown thus - - - - r = component number.

(*a*) Sine-wave modulation.
(*b*) Square-wave modulation.
(*c*) Trapezoidal-wave modulation, $s = 0 \cdot 05$.
(*d*) Trapezoidal-wave modulation, $s = 0 \cdot 11$.
(*e*) Trapezoidal-wave modulation, $s = 0 \cdot 18$.
(*f*) Trapezoidal-wave modulation, $s = 0 \cdot 31$.

and tests were therefore made at modulation indices of 1 and 3, to compare the spectra of signals using the two waveforms and having the same total build-up time. The waveform used for the half-sine-wave test was only an approximation, formed by clipping the corners of a trapezoid, as shown in Fig. 3(*b*), but a comparison was made of the measured results with the theoretical envelope of the spectrum using true half-sine-wave build-up, and the agreement was found to be within ± 2 dB. The value of s of $0 \cdot 15$ used in the half-sine-wave build-up tests was somewhat different from those used in the earlier tests with trapezoidal waveforms. Appropriate trapezoidal-modulation envelope values were therefore obtained by interpolation from the measured values, a process which is justified by the smooth way in which the envelope shifts with changing values of s. In the region where the envelope amplitudes are about -60 dB relative to carrier level the measured envelope for half-sine-wave build-up was about 6 dB higher than that for straight-line build-up of the same s, the difference becoming progressively smaller for frequencies nearer the carrier. A comparison was also made between the spectra for half-sine-wave build-up and straight-line build-up having the same build-up rate as the centre portion of the half-sine-wave approximation. The effect of clipping the corners of the straight-line build-up was generally less than 2 dB. Thus it seems that the second-order discontinuities at the corners of a trapezoidal modulating wave do not, of themselves, contribute significantly to the bandwidth.

Tests were next made with the waveforms of Figs. 3(*c*) and 3(*d*) to determine whether different parts of the build-up waveform contributed more than others to the spread in the spectrum. In Fig. 3(*c*) it will be seen that the middle 50% of the transition has infinite slope whereas the build-up time for the extremes of the transition, corresponding to $s = 0 \cdot 14$, may be assumed to contribute little to the skirts of the spectrum. Similarly, in Fig. 3(*d*) the first and last 25% of the transition have infinite slope whereas the middle has a build-up ratio of $0 \cdot 14$. The spectra due to the waveforms of Figs. 3(*c*) and 3(*d*) were found to differ considerably in shape. With the waveform of Fig. 3(*c*) the spectrum was similar to that produced by rectangular modulation, alternate components forming the envelope, which decreased uniformly in amplitude with increase in difference of frequency from the carrier. With the waveform of Fig. 3(*d*) the envelope appeared to be formed by clusters of components separated by intervals of approximately twelve components; the sensitivity of the spectrum analyser was, however, insufficient to detect more than one cycle of amplitude variation on each side of the carrier. The envelopes due to the two waveforms were of roughly the same amplitude, so that it seems reasonable to assume that all parts of a transition having an equal slope contribute a similar amount to the spectrum. The envelope due to Fig. 3(*c*) was 5 dB lower at its skirts than that due to a rectangular waveform when $m = 1$, and 8 dB lower when $m = 3$. Comparison of the spectrum of Fig. 3(*c*) with that due to a

Table 1

ENVELOPE OF SPECTRUM DUE TO TRAPEZOIDAL MODULATION

m	1				2				3					3·01	6				
	s				*s*				*s*				C.C.I.R. limit	C.C.I.R. limit	*s*				C.C.I.R. limit
r	0·05	0·11	0·18	0·31	0·05	0·11	0·18	0·31	0·05	0·11	0·18	0·31			0·05	0·11	0·18	0·31	
0	−4	−4	−4	−4	−42	−42	−44	−38	−12	−12	−12	−10	0	0	−47	−43	−32	−19	0
1	−6	−6	−6	−6	−7	−7	−7	−7					0	0	−20	−18	−16	−16	0
2	−12	−12	−12	−14	−6	−6	−6	−6	−6	−8	−8	−6	0	0					0
3					−11	−11	−11	−12	−6	−6	−6	−7	0	0	−17	−15	−15	−14	0
4	−28	−28	−29	−34					−12	−11	−10	−13	0	0					0
5					−24	−24	−26	−30					−19	−21	−8	−8	−8	−8	0
6	−35	−37	−40						−23	−25	−26	−30	−24	−27	−6	−6	−6	−7	0
7				−54	−30	−32	−37						−27	−33	−10	−10	−10	−12	0
8	−40	−44	−50	−56				−44	−30	−31	−36		−31	−38					0
9					−36	−38	−48				−42	−42	−33	−42	−22	−23	−27		−19
10	−45	−48		−59					−35	−38			−36	−46			−32	−30	−23
11					−39	−44		−54			−47	−51	−38	−49	−27	−30			−26
12	−48	−55					−54		−38	−45		−52	−40	−52			−37	−40	−29
13					−42	−50					−51		−42	−55	−30	−36			−32
14	−52	−59							−42	−50			−44	−58			−42		−35
15					−45	−56	−59				−59		−46	−60	−35	−43		−47	−37
16	−53								−45	−59			−48				−46	−53	−40
17					−50	−59							−49		−39				−43
18	−56								−48				−51			−50		−55	−45
19					−52								−52		−41		−56		−47
20	−58								−51				−53			−51			−49
21					−53								−55		−44		−58	−60	−50
22	−59								−54				−56			−54			−52
23					−56								−57		−46				−54
24									−55				−58			−56			−55
25					−58								−59		−49				−57
26									−56				−60			−59			−58
27					−58										−50				−59
28									−60										−60
29															−53				
30																			
31															−55				
32																			
33															−58				
34																			
35															−60				

The levels are those of the envelope-forming components only, expressed in decibels relative to the unmodulated carrier.

trapezoidal wave having the same overall build-up time showed that the envelope due to the trapezoidal wave became increasingly lower with increasing difference of frequency from the carrier.

The results of the work described in this Section may be summarized by stating that of a number of modulating waves having different build-up waveforms the narrowest spectrum for a given build-up time and a given modulation index over the range tested was that produced with straight-line build-up, i.e. by a trapezoidal waveform. A rough test indicated that all parts of the transition between mark and space frequencies having the same build-up rate contribute approximately equal amounts to the skirts of the spectrum. It was concluded that, for the purpose of f.m. telegraphy transmission, nothing is to be gained from the use of anything other than a simple trapezoidal modulating waveform and that it was unnecessary to continue the tests on other waveforms.

(6) DISCUSSION ON SPECTRA DUE TO TRAPEZOIDAL MODULATION

As a matter of interest, the results for trapezoidal modulation have been compared with those presented by Japan to the VIIth Plenary Assembly of the C.C.I.R. The spectra in the Japanese document are measured spectra for the case of trapezoidal modulation with a build-up ratio of 0·1 and modulation indices of 3, 6 and 8. For $m = 3$, the spectrum calculated from an approximate formula is also given. The most striking difference between the two sets of results is that, whereas for

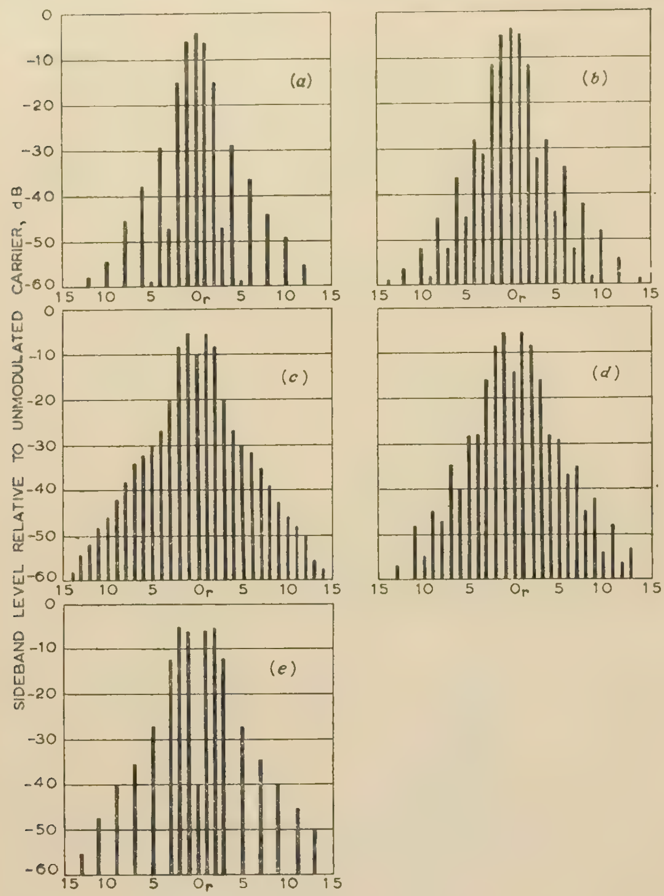


Fig. 5.—Typical measured spectra of f.m. signal with trapezoidal waveform modulation, $s = 0\cdot11$.

r = Component number.

(*a*) $m = 1\cdot0$. (*b*) $m = 1\cdot2$. (*c*) $m = 1\cdot5$. (*d*) $m = 1\cdot8$. (*e*) $m = 2\cdot0$.

$s = 0\cdot11$ alternate components were found to form the envelope out to beyond the 16th component for $m = 3$ and out to the 15th component for $m = 6$ (see Table 1), this is only so in the Japanese measurements ($s = 0\cdot10$) out to the 6th component for $m = 3$ and the 7th component for $m = 6$. In the Japanese calculated spectrum for $m = 3$, alternate components out to the 8th component form the envelope. The reason for the differences between the two sets of results is not understood. Nevertheless, if the envelopes of the Japanese spectra are compared with those of Table 1, interpolated for $s = 0\cdot1$, the agreement is within ±5 dB, the latter results being generally the higher for $m = 3$ and the lower for $m = 6$.

The desirable limits to the emitted spectra of F1 transmissions are defined by the C.C.I.R.[4] for modulation indices between 2·5 and 20. The necessarily occupied bandwidth is defined for modulation indices between 2·5 and 8 as $2\cdot5 \times$ (deviation) $+ 0\cdot5 \times$ (telegraph speed). The spectrum of the out-of-band radiation should not exceed -15 dB at the limits of the necessarily occupied bandwidth and should fall to -60 dB by at least 17 dB per octave for $2\cdot5 < m \leqslant 3$ and 25 dB per octave for $3 < m \leqslant 8$. The C.C.I.R. limits for $m = 3$ and $m = 6$ have been included in Table 1. Since the slope changes from 17 to 25 dB per octave at $m = 3$, the C.C.I.R. limits for both $m = 3$ and $m = 3\cdot01$ are given in the Table. The limits are stated to be just met by a signal shaped by a suitable filter to have a build-up time between 0·1 and 0·9 of the steady-state value equal to 8% of one telegraph element. Such a signal would probably have an overall build-up ratio of about 0·14. It will be seen that by employing a trapezoidal signal a build-up ratio of 0·11 may be employed for $m = 3$ and $m = 6$ whilst still retaining the spectrum within the C.C.I.R. limits. A build-up ratio of 0·31 is required to meet the C.C.I.R. limit for $m = 3\cdot01$ which does not permit of an 8% build-up time.

As previously discussed, for the case of half-sine-wave build-up it is possible to normalize the spectrum envelopes with respect to m so as to produce a universal set of curves, as in Fig. 1(*a*). An attempt has been made to do the same thing empirically, using the tabulated data, for trapezoidal modulation. The result appears in Fig. 1(*b*). When checked against the original measurements, the normalized curves gave results generally accurate within ±3 dB, but occasionally in error by up to ±5 dB. Whilst it has not been formally proved that normalization is justified for trapezoidal modulation, it is thought that the curves give a useful degree of accuracy, should it be desired to interpolate the measurements over the range of the tests. A comparison of Figs. 1(*a*) and 1(*b*) confirms that the bandwidth with trapezoidal modulation is less than that with half-sine-wave build-up for the same total build-up time, the difference being most apparent for $ms = 0\cdot1$–$0\cdot5$ and tending to zero for $ms = 2$ and, of course, $ms = 0$.

(7) CONCLUSIONS

Experimental evidence suggests that for low values of modulation index and for total build-up times of about 0·1 of a telegraph element, a straight-line transition between mark and space frequencies, i.e. trapezoidal modulation, in frequency-shift telegraphy gives the minimum bandwidth. The waveform of such a transition has the considerable added advantage that it is very easily generated. If integral values of modulation index are used it is found that lower-order spectrum components having numbers in the series $m \pm 2$, $m \pm 4$ . . . disappear, progressively higher-order components showing the same tendency as the build-up time is reduced. Although the primary purpose of the work was to obtain information on the best keying waveform for radiotelegraph circuits, the results apply equally to any frequency-modulated signals.

(8) ACKNOWLEDGMENTS

The authors wish to acknowledge assistance from several of their colleagues and, in particular, to thank Mr. H. B. Law for his helpful guidance and Mr. D. L. Lisney for most of the experimental work. Acknowledgment is also made to the Engineer-in-Chief of the Post Office and to the Controller of Her Majesty's Stationery Office for permission to publish the paper.

(9) REFERENCES

(1) C.C.I.R. VIIth Plenary Assembly, London, 1953. Study Programme No. 39(I).
(2) C.C.I.R. VIIth Plenary Assembly, London, 1953. Study Programme No. 43(II).
(3) CAWTHRA, W. A., and THOMSON, W. E.: 'Bandwidth of a Sinusoidal Carrier Wave, Frequency-Modulated by a Rectangular Wave with Half-Sine-Wave Build-up', *Proceedings I.E.E.*, Paper No. 1066 R, January, 1951 (**98**, Part III, p. 69).
(4) C.C.I.R. VIIth Plenary Assembly, London, 1953. Recommendation No. 87, Bandwidth of Emissions.
(5) PARKER, P. N., and GREGORY, G.: 'Banbury Radio Measuring Station', *Post Office Electrical Engineers' Journal*, October, 1954, **47**, p. 143.

[The discussion on the above paper will be found on page 147.]

621.396.812.3

Paper No. 2150 R
Aug. 1956

AN IMPROVED FADING MACHINE

By H. B. LAW, B.Sc.Tech., F. J. LEE, R. C. LOOSER, B.Sc.(Eng.), Associate Members, and F. A. W. LEVETT.

(*The paper was first received* 27*th February, and in revised form* 14*th May*, 1956. *It was published in August*, 1956, *and was read before the* RADIO AND TELECOMMUNICATION SECTION 14*th November*, 1956.)

SUMMARY

A fading machine which formerly gave cyclic fading has been modified to give substantially random fading with the amplitude of the signal conforming to the Rayleigh distribution, as with long-distance h.f. radio signals. The fading is obtained by combining six components in random phase. The machine simulates propagation by up to three paths with a spread of up to 2 millisec in path-time delay and caters for dual diversity reception. It has proved to be a valuable tool in laboratory investigations of radiotelegraph equipment.

LIST OF SYMBOLS

f_0 = Centre frequency of fading-signal power spectrum.
l = Characteristic size of ionospheric irregularity.
N = Quasi-frequency of fading.
$N(p)$ = Frequency of upward crossing of power level p.
P = Mean power of fading signal.
p = Instantaneous power.
$\mathrm{P}(p)$ = Probability that signal power is less than p.
T^2 = Variance of time autocorrelation function.
u_w = Velocity of ionospheric drift.
V = R.M.S. signal amplitude.
v = Instantaneous signal amplitude.
$W(f)$ = Power spectrum.
$\rho(t)$ = Time autocorrelation function.
σ^2 = Variance of power spectrum.

(1) INTRODUCTION

Equipment for long-distance radiocommunication in the 4–30 Mc/s frequency band is judged by its performance in combating, in the presence of noise and interference, the fading and multi-path propagation phenomena characteristic of ionospheric propagation. Tests in service are apt to be slow and inconvenient, and a means of simulating the natural phenomena under controlled conditions is very desirable. A fading machine for producing frequency-selective fading effects by combining two or three differently delayed signals of constant amplitude but varying relative phase was described by Bray, Lillicrap and Owen[1] in 1947. It is now known,[2] however, that natural fading is very different from the cyclic variations of signal strength produced by the original fading machine in the two-path-propagation condition in which it was normally used. The difference was probably of little consequence in subjective comparisons of various double-sideband and single-sideband radiotelephone systems for which the machine was first used. It was, however, more important when the machine was applied in radiotelegraphy development and testing, and it was desired to compare performances objectively. For the results of such work to be really useful the test conditions must simulate the conditions encountered in practice with sufficient accuracy, and this requires, in the first place, a knowledge of the fading, multi-path, noise and interference effects that occur on practical circuits. Although much remains to be learned on these topics, the knowledge of fading has advanced sufficiently, not only to make the deficiencies of the old fading machine obvious, but also to indicate a simple way of improving it. The paper describes a modified fading machine giving substantially random fading similar to that observed in natural propagation.

Mr. Law, Mr. Lee and Mr. Levett are at the Post Office Research Station.
Mr. Looser is in the Post Office Engineering Department, Radio Planning Branch.

(2) CHARACTERISTICS OF NATURAL FADING

The fading of a signal can be described in terms of the distribution in time of the signal amplitude and of the rapidity of fading. So far as the amplitude is concerned, use is commonly made of the Rayleigh distribution in fading studies,[2,3] and in the authors' experience long-distance h.f. signals conform closely to this distribution, which may be written in the form

$$\mathrm{P}(p) = 1 - \exp(-p/P) \quad . \quad . \quad . \quad . \quad (1)$$

where P is the mean signal power and $\mathrm{P}(p)$ is the proportion of time for which the signal power is less than p. It is necessary to add a proviso that the time interval of observation should not be so long that appreciable variations can occur in the mean power, due for example to changes of ionospheric absorption or focusing; in practice, periods of the order of five minutes generally satisfy this proviso. Conformity of the signal strength to the Rayleigh distribution is consistent with the theory that the signal received is the resultant arising from the combination in random phase of a large number of signal components. The fading signal may be considered[2] equivalent to the result of passing white Gaussian noise through a filter having a narrow pass band centred on the carrier frequency.

The speed of fading can conveniently be expressed in terms of the frequency at which the signal amplitude increases through some specified value. It is shown in Section 10.1, by using a result given by Rice,[4] that, considered as a function of amplitude, the frequency of upward crossing is at a maximum at the most probable amplitude or, in terms of power, at a level equal to half the mean power. This maximum frequency may be called N and defined as the quasi-frequency of fading. It is convenient to count crossings of the half-mean-power level in practical measurements of frequency of fading, for, the frequency being at a maximum, the measurement is insensitive to errors in defining the reference level; also, since the level is reasonably high, the probability of error due to interference by unwanted signals is comparatively small. Rewriting eqn. (11) of Section 10.1 in terms of power, we have for the relationship between the frequency of upward crossing of a power p, $N(p)$ say, and the quasi-frequency N

$$N(p)/N = (2p/P)^{1/2} \exp(\tfrac{1}{2} - p/P) \quad . \quad . \quad . \quad (2)$$

Observations on long-distance h.f. signals in the 10–20 Mc/s band have given results in good agreement with eqn. (2); values of N have generally been in the range 4–15 fades per minute, though more rapid flutter fading has been observed on occasion. These figures may be compared with median values of fading frequency of 28, 12 and 9 fades per minute given[5] for transatlantic

signals received in Germany on 10, 15 and 20 Mc/s respectively; a reference level about 6 dB lower than that used above to define the quasi-frequency of fading was used in the German measurements and the effect of this would be to reduce the frequency as measured, relative to the frequency according to the definition given above, in the ratio 3 : 2. Thus the frequencies determined by the authors are rather low compared with the more extensive German results. Another source of data lies in studies of fading of vertically incident signals. These studies have shown[6,7] that fading is mainly due to the drift of a relatively unchanging diffraction pattern past the receiving aerial, the drift being caused by horizontal movement in the ionosphere. Ionospheric drift velocities of the order of 50–100 m/s are normal[8] though velocities may reach 1 000 m/s in the F-region under disturbed conditions. The diffraction pattern arises from irregularities in the ionosphere, and a characteristic size of irregularity of 200 m has been determined.[7] By means of a relationship between the time autocorrelation function and the power spectrum given by Booker, Ratcliffe and Shinn[9] and results given by Rice[4] relating the power spectrum to the quasi-frequency, it is shown in Section 10.1 that the quasi-frequency is given by

$$N = 0 \cdot 29\, u_w / l \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where u_w is the velocity of horizontal drift of the ionospheric layer and l is the characteristic size of irregularity. Inserting typical values we have

$$\begin{aligned} N &= 0 \cdot 29\ (100/200) \\ &\simeq 0 \cdot 15\,\mathrm{sec}^{-1} \\ &= 9\ \text{fades/min} \end{aligned}$$

This is for vertical incidence. To get an idea of the effect at oblique incidence we assume with McNicol[3] that the diffraction pattern on the ground is coarser than that for vertical incidence by a factor $\sec\theta$, where θ is the angle of incidence of the waves at the ionosphere. Thus, taking 70° as a typical value for θ for long-distance signals we have

$$\begin{aligned} N &= 9 \cos 70^\circ \\ &= 3\ \text{fades/min} \end{aligned}$$

This is slower than the fading frequencies observed by the authors, but there are some factors, of which no account has been taken, which would tend to increase the speed of fading. Thus on long-distance h.f. circuits we are generally concerned not with one but with several ionospheric reflections; also ground or sea reflections will introduce additional scatter. It is to be expected that the fading-frequency effects of a number of reflections will add up on an r.m.s. basis. It may be concluded that the result given by applying the very extensive data for vertical incidence is not inconsistent with the direct measurements at oblique incidence.

Multi-path propagation may be described in terms of the number of propagation paths, their relative activities and their relative path-time delays. Little new information on the long-distance h.f. case seems to have been published since the data were summarized in the paper on the original fading machine.[1] Measurements on pictures transmitted by radio, made by the authors' colleagues, J. W. Allnatt and E. D. J. Jones, have shown that the path-time-delay spread on some typical working circuits rarely exceeded 2 millisec during three years including the recent sunspot minimum; spreads in the range 0–1·5 millisec were common.

(3) DESIGN CONSIDERATIONS

Evidently an artificial fading machine should produce Rayleigh fading with a quasi-frequency in the range 4–40 fades/min and should cater for multi-path propagation with a path-time-delay spread adjustable up to 2 millisec. Since the performance of radio receivers usually deteriorates as the speed of fading increases, it is desirable to have available for test purposes fading quasi-frequencies in the middle and upper parts of the range of frequencies commonly observed. Obviously the machine should include facilities for the injection of noise and for the measurement of signal/noise ratio. It should provide a transmission path suitable for the usual types of signal used in long-distance radiotelegraphy and preferably also for commercial-quality telephony. It should cater for dual spaced-aerial diversity reception. In the present case it was desired, as a matter of practical convenience, to use such parts of the old fading machine as could appropriately be fitted into the new one; this applied particularly to the 100 kc/s channel filters and, above all, to the elaborate audio-frequency delay units. The old machine, in fact, satisfied all the main requirements apart from that of producing Rayleigh fading, so that the problem was essentially that of modifying it to overcome this limitation.

The most obvious way of producing Rayleigh fading would be to modulate signals in amplitude with a control signal varying according to the Rayleigh distribution at a suitable speed. This arrangement is, however, open to the objection that the fading would not be accompanied by phase changes. The combination of many components in random phase, which is the cause of fading in nature, causes marked swinging of the phase of the resultant, and, since this is likely to have significant effects in radiotelegraph receiving equipment, particularly in frequency-modulated systems, it is important that the fading machine should simulate the phase changes as well as the amplitude variations.

Since the combination of many components in random phase gives rise to the Rayleigh distribution, a carrier varying in exactly the desired manner, in both amplitude and phase, could be obtained by selecting a narrow band of noise by means of a band-pass filter. The bandwidth of the filter would have to be a fraction of a cycle per second in the present case to obtain the desired frequency of fading. To select such a band of noise directly at the 100 kc/s intermediate frequency used in the fading machine would be difficult; if the required bandwidth were obtained at a much lower frequency, which would be easier, additional equipment would be needed to translate the fading carrier to the 100 kc/s region. Also, however the fading carrier were obtained, the output of the modulator used to mix it with the signals would have to vary linearly, not only with the amplitude of the carrier, as is obvious, but also with the signal amplitude, in order to avoid distortion of the envelope of amplitude-modulated signals.

As there seemed likely to be serious practical difficulties in the application of the filtered-noise method, in which, in effect, an infinite number of components is used, the question arose whether a sufficiently good approximation to the Rayleigh distribution could be obtained with a fairly small number of carrier components, which could be individually modulated by the signals with combination after modulation. The probability distributions for small numbers of oscillations combined in random phase were studied by Miss Slack.[10] She was mainly interested in determining the probability of large amplitudes, being concerned with the loading of multi-channel transmission systems, and her results, given graphically, are unsuitable for use at the small amplitudes that are of primary interest in fading studies. It was therefore necessary to go back to the Tables prepared by Pearson[11] about fifty years ago. He studied a problem of random migration in which the migrant makes n hops, each of length a, the directions of the hops being random in two dimensions. His Tables show the probability of a

migrant landing in an elementary area dA distant r from the starting-point for values of n from 2 to 7. Evidently, the probability of a migrant landing between r and $r + dr$ from the starting-point is $2\pi r$ times the values given in the Tables. By multiplying in this way and then integrating, the probability of a migrant reaching a distance greater than r from the starting-point can be determined. These processes have been applied to Pearson's figures for $n = 3$, 4, 5 and 6, with the results shown in Fig. 1, in which the amplitudes are plotted on an arbitrary

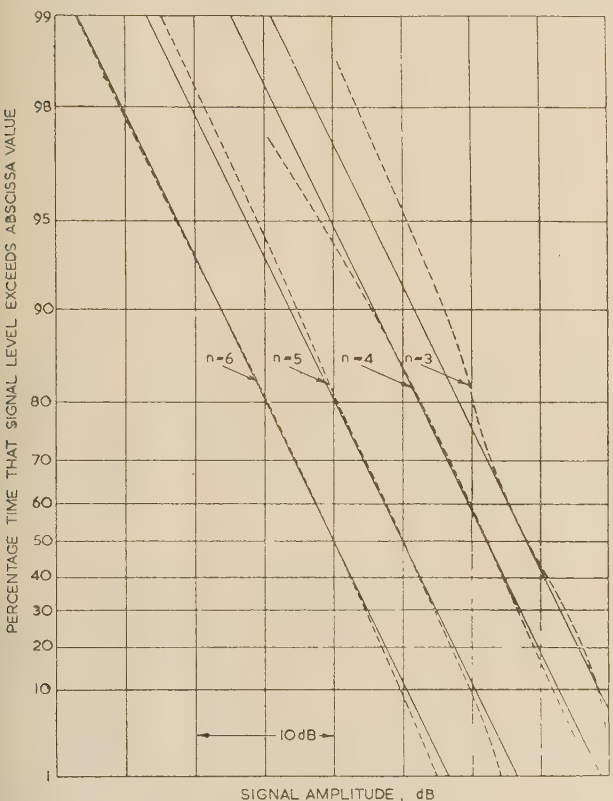


Fig. 1.—Calculated amplitude distributions for n vectors (dashed curves) compared with Rayleigh distribution (straight lines).

decibel scale; each straight line represents the Rayleigh distribution having the same median value as the associated approximation. The fit of the approximate distributions improves rapidly as n is increased and is good when $n = 6$. It was therefore decided to rearrange the fading machine so that each output would be composed of six randomly phased components.

The speed of fading in the revised machine would be dependent on the frequency spread of the component carriers. A rough idea of the relationship may be gained by supposing that the fading arises from beating between two groups of carriers, corresponding to the upper and lower halves of the carrier-frequency spectrum. It is to be expected that, with approximately uniform spacing of the carrier frequencies, the quasi-frequency of fading will be about half the carrier-frequency spread, which would thus have to be of the order of 0·5 c/s to produce a fading quasi-frequency of 20 fades/min. The frequency instability of the carrier supplies should be small enough to maintain reasonable separation between the frequencies of adjacent oscillators, to avoid very slow beats, but large enough to avoid recurrent fading patterns due to the maintenance of some simple relationship between the carrier frequencies for an appreciable part of a test period. It was envisaged that test periods in radiotelegraph work would range from a few minutes to a quarter of an hour, the latter interval corresponding to the transmission of about 90 000 telegraph elements at speeds, common in machine radiotelegraphy, of the order of 100 bauds. With carrier-frequency separations of about 0·1 c/s, minute-to-minute frequency variations of the order of 0·01 c/s seemed appropriate; this order of stability is attainable in simple 100 kc/s quartz-crystal oscillators.

(4) DESCRIPTION OF THE MACHINE

The fading machine can be set up in various ways to suit different requirements, but it is convenient to describe it as it is normally used in tests on radiotelegraph equipment, with brief comments on alternative arrangements. The normal arrangement, shown in Fig. 2, simulates dual-space-diversity reception of signals arriving by two radio paths having different propagation times. The input, in the range 100–6 000 c/s, is split, and one part goes through a delay unit of the low-pass-filter type, adjustable in steps of 0·03 millisec over the range 0–2 millisec, to simulate the difference of path-time delay; a second adjustable delay unit is available to cater for three-path-propagation tests. Next, the audio-frequency signals are taken into resistive splitter pads, each giving six equal outputs, which are then combined in the random fading units. For the two-path condition depicted in Fig. 2 each random fading unit receives three inputs from the splitter pad that is fed with delayed signals and three non-delayed inputs.

A random-fading unit (Fig. 3) consists of an assemblage of six balanced modulators each of which is fed with carrier, nominally at 100 kc/s, from a separate quartz-crystal oscillator. The modulator outputs are combined in a resistive network, and, by arranging that the oscillator frequencies are suitably spread over a small frequency-band and have suitable stabilities,

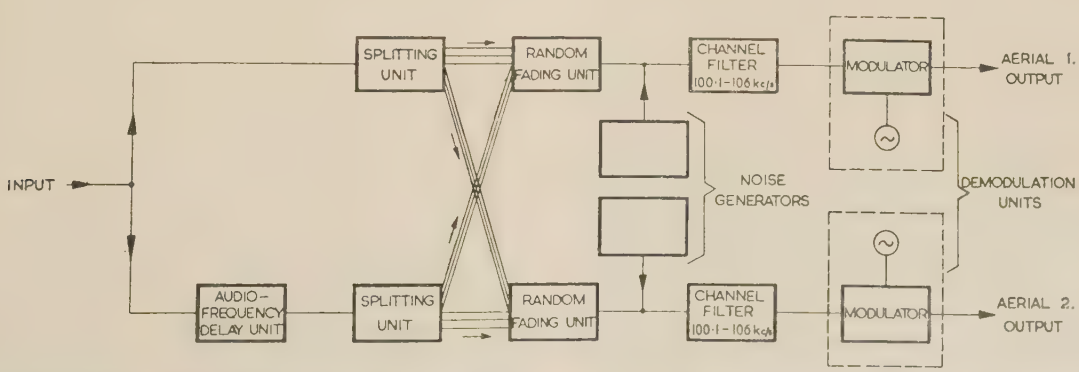


Fig. 2.—Arrangement for simulating spaced-aerial-diversity reception of fading signals under two-path propagation conditions

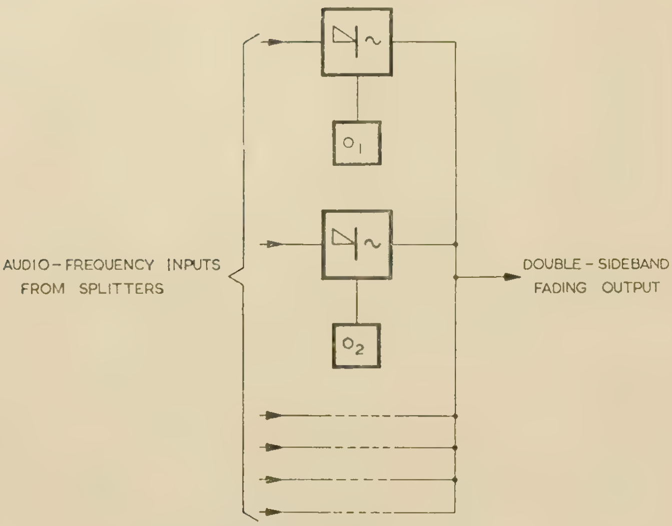


Fig. 3.—Random-fading unit.

the modulator outputs combine in substantially random phase and the resultant is a signal fading at the desired speed. The oscillators are simple units employing $+5°$ X-cut quartz crystals without temperature control; variable capacitors associated with the crystals cater for frequency adjustment. Some difficulty was experienced in the early stages of the work from a tendency for the frequency of one oscillator to slide through that of another as the temperature of the equipment changed, causing slow cyclic changes in the effective signal strength in the process. This trouble was cured by distributing the oscillators in frequency in the order of their temperature coefficients, so that adjacent oscillators do not have widely different temperature characteristics. Facilities are provided for recording the signal level so that slow beats may be detected. Also, oscillator frequencies may be compared in pairs in terms of the speed of rotation of a rotary phase-meter, which shows the relative phase of the two oscillator outputs; any pair of oscillators can be selected for comparison by multi-way switches without disturbance to tests in progress. Entirely separate batches of oscillators are used for the two random-fading units so that the fadings in the two diversity paths are uncorrelated.

The outputs of the fading units are filtered in band-pass filters (100·1–106 kc/s) to remove the lower sideband and residual carrier, and the signals are then demodulated, using a local carrier, to restore them to the original audio-frequency range. Equipment trials often call for signals at radio or intermediate frequencies, and in such cases it is necessary to translate the audio-frequency outputs of the machine; in this connection facilities for extracting outputs at the fading-machine intermediate frequency are sometimes useful.

Each diversity path has associated with it a noise source for injecting into the i.f. path Gaussian noise uniformly distributed over the range 90–110 kc/s. The noise is derived from a gas-discharge tube via an amplifier provided with automatic amplitude control. Attenuators provide for the adjustment of signal and noise levels, which are measured by means of a thermal meter, the noise bandwidth being 6 kc/s, defined by the i.f. filter. Signal levels are normally set up with the machine in the non-fading condition, only one of the six signal components being connected through. Since the components are normally equal, the mean signal power in the fading condition is simply six times the power of the single component.

The arrangement described corresponds to a two-path condition in which the two paths are equally active, i.e. they produce the same mean signal power at the receiver. This is a peculiar condition, which is unlikely to be maintained for long in natural propagation, because the mean powers of signals arriving by different routes usually vary slowly but independently, owing perhaps to changing absorption and focusing effects in the different ionospheric regions involved. The condition of unequal path activity may be simulated on the machine by attenuating the signal in, for example, the delayed path. If the attenuation is heavy the resultant signal approximates to a three-component one and the approximation to Rayleigh fading is comparatively poor. In these circumstances it is better to reduce the attenuated path to a single component, leaving five normal-sized components for the stronger path; the fading still gives a good approximation to the Rayleigh distribution, and, although the weaker-path signal does not fade at all, the probability of its exceeding the stronger-path signal, which is usually the point of prime importance in two-path radiotelegraphy studies, is substantially correct, as is shown in Section 10.2. Due allowance is, of course, necessary for any attenuated signal components in deriving the total mean power from the power of a single component of normal amplitude.

In addition to catering for radiotelegraphy tests the normal arrangement of the fading machine is immediately applicable to single-sideband telephony work. A facility is provided for upsetting the carrier balance of the modulators in the fading units so that a double-sideband signal results, with full carrier; by omitting the sideband filters and substituting envelope detectors for the normal demodulation units, double-sideband telephony tests are possible.

(5) MEASURED FADING PERFORMANCE

The amplitude distribution of signals from the fading machine has been determined by means of a level-distribution analyser, which determines the proportions of time, during a test period, for which various predetermined levels are exceeded. A typical result is shown in Fig. 4; the plotted points represent analyser readings and the straight line is the Rayleigh distribution. The points are a good fit to the straight line at low and medium levels, falling away somewhat at the higher levels. The difference between the measured and Rayleigh distributions at the higher levels is partly attributable to the small number of signal components, as reference to the curve for $n = 6$ in Fig. 1 shows. The discrepancy is of no importance in practice because it is the low signal levels that cause trouble in radio reception, and the machine does produce these in the correct proportions.

The level-distribution analyser incorporates counters which indicate the number of crossings of the various levels to which it has been set, and the points in Fig. 5 show the relative frequencies of crossings plotted against signal level. The curve is the theoretical relationship of eqn. (2); it is to be noted that the signal level is expressed relative to the median level, which is 1·6 dB below the r.m.s. amplitude or mean power. Here again, the fit of the measured points to the theoretical curve is good except at the higher signal levels. With a frequency spacing of 0·1 c/s between oscillators a quasi-frequency of fading of 19 fades/min was obtained. It has been found in practice that this is about the lowest speed of fading that the machine can give without serious risk of slow beats between oscillators. The maximum total spread of oscillator frequency is 12 c/s, giving a fading quasi-frequency of about 450 fades/min.

The level-distribution analyser gives no information on the variation of phase angle of the resultant signal, and, as has already been remarked, this variation is of some importance in radiotelegraphy. In order to throw some light on this point an experiment was performed in which a fading signal was resolved

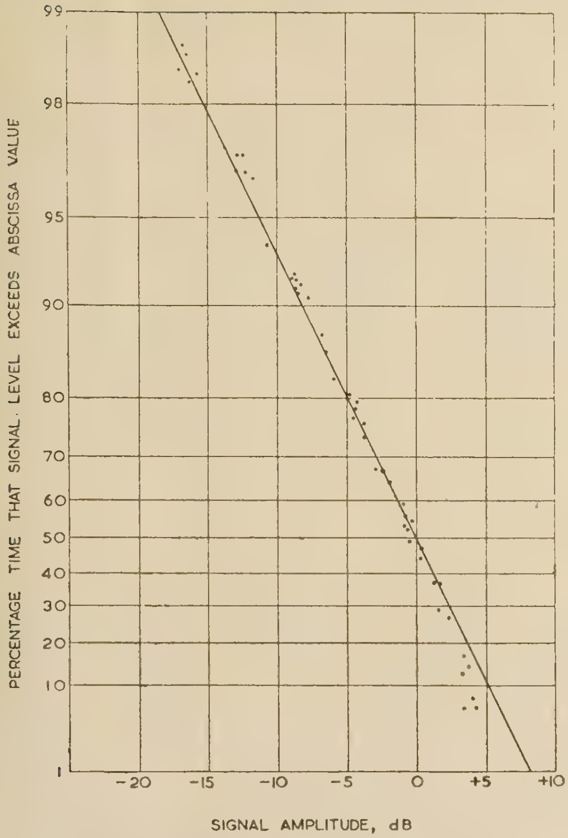


Fig. 4.—Measured amplitude distribution.
The straight line is the Rayleigh distribution.

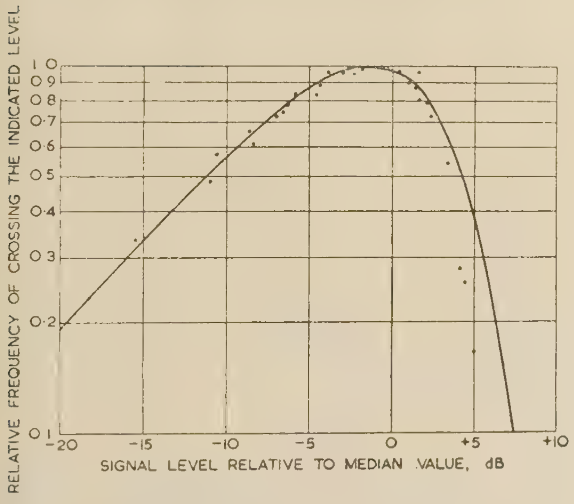


Fig. 5.—Variation of frequency of fading with depth of fade.
Fading-machine results compared with theoretical curve.

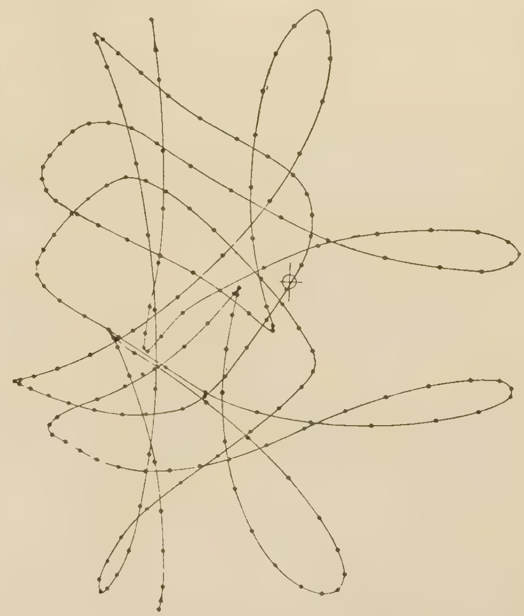


Fig. 6.—Locus of fading-signal vector from fading machine.
25 sec exposure. Dots represent 8 c/s timing wave.

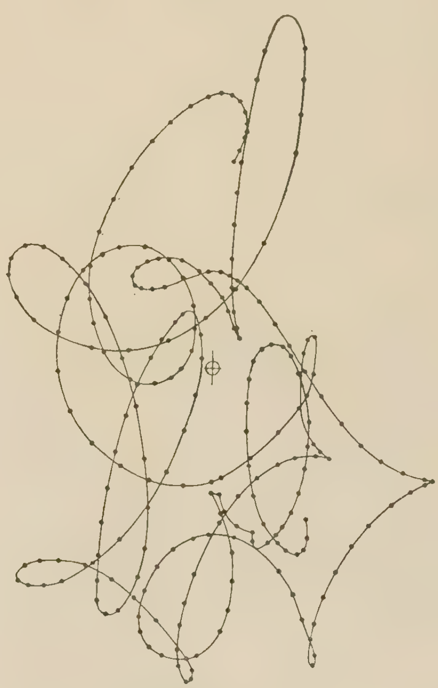


Fig. 7.—Locus of fading-signal vector from WWV, 15 Mc/s, on June 28th, 1955.
30 sec exposure. Dots represent 10-c/s timing wave.

into sine and cosine components, by reference to a local source, and the amplitudes of the components were arranged to control respectively the X and Y deflections of the beam of a cathode-ray tube; the position of the spot relative to the origin, given by the no-signal condition, thus gave the amplitude and phase of the signal at any instant. By modulating the spot brightness with a timing wave and taking a long-exposure photograph of the display, a picture of the evolutions of the resultant signal vector over the period of the exposure was obtained. Fig. 6 shows a typical result obtained with a signal passing through the fading machine, and Fig. 7 shows a result obtained with a radio signal received in England from the standard-frequency transmitter WWV, Washington, on 15 Mc/s. It was necessary to use a standard-frequency signal for this purpose to permit proper synchronization of the local source in terms of which the signal was resolved.

(6) EXPERIENCE IN USE AND POSSIBLE IMPROVEMENTS

The improved fading machine has been used as a tool in several h.f.-radiotelegraphy investigations, but detailed reports

on these would be inappropriate in the present paper. The change from the old machine to the new one has, however, brought out some points of general interest which can perhaps usefully be discussed here. In the first place, the use of the new machine has led to a striking change in the appreciation of the performance of systems on test. The use of cyclic fading in the old machine, often with rather deep fades, e.g. a maximum-to-minimum range of 20 or 30 dB, tended to direct the attention of the experimenter to trivial details of system behaviour when the system was subjected to conditions that are rare in nature. The result was therefore apt to be misleading and, in addition, the mass of detail was difficult to appreciate. With the close approximation to natural fading given by the new machine the various signal levels are automatically produced in the correct proportion and the danger of false emphasis does not arise. Furthermore, the use of the new machine in tests on f.m.-radio-telegraphy systems has helped to establish a succinct method[12] of describing noisy-signal performance in terms of regenerated-signal error liability; this has made appreciation of results much easier.

The acid test of a simulating device is that the results obtained by its means should agree sufficiently well with those obtained under practical conditions. The new fading machine does not so far fully pass this test.[13] The bulk of the work done with it has been with white Gaussian noise as the sole disturbance accompanying the signals, but it is known that on practical h.f. channels unwanted signals, man-made interference and discrete atmospheric crashes are usually more disturbing than the white background noise. It is, of course, possible to inject any kind of disturbance into the signal path at the output of the fading machine, and typical atmospheric crashes, reproduced from a magnetic tape record, have occasionally been used in this way. It is hoped to provide a regular facility on these lines when studies of atmospherics have led to an adequate way of specifying their characteristics in terms of fundamental measurements. A regular facility for injecting interference would be a worth-while addition also.

In its present form the fading machine can simulate propagation by up to three independent ionospheric paths. It is known that more than three paths may sometimes be active on practical radio channels, but whether this is of sufficient importance in its effects to make the present limitation of the fading machine a serious one is not known. The practice of using a working frequency fairly close to the maximum usable frequency, so that the number of possible modes is limited, supports the hope that the present arrangement will suffice. Similarly path-time-delay spreads much larger than the 2 millisec spread available in the fading machine are occasionally reported, but here again the working-frequency argument, supported this time by echo measurements on pictures sent by radiotelegraph, suggests that the delay range is adequate.

(7) CONCLUSIONS

Signal-level distribution analyses indicate that, as would be expected from theory, the combination of six equal signal components in random phase gives a good approximation to the Rayleigh distribution, which is characteristic of the fading of long-distance h.f. radio signals. The fading machine would therefore appear to be adequate—as a fading machine. To complete the laboratory simulation of practical h.f. radio channels an adequate source of atmospheric noise and interference is now needed.

(8) ACKNOWLEDGMENTS

Acknowledgment is made to the Engineer-in-Chief of the Post Office and to the Controller of Her Majesty's Stationery Office for permission to publish the paper. The authors express their thanks to Mr. W. A. McClure for the integration of Pearson's tables.

(9) REFERENCES

(1) Bray, W. J., Lillicrap, H. G., and Owen, F. C.: 'The Fading Machine, and its Use for the Investigation of the Effects of Frequency-Selective Fading,' *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 283.

(2) Ratcliffe, J. A.: 'Diffraction from the Ionosphere and the Fading of Radio Waves', *Nature*, 1948, **162**, p. 41C5.

(3) McNicol, R. W. E.: 'The Fading of Radio Waves of Medium and High Frequencies', *Proceedings I.E.E.*, Paper No. 891 R, November, 1949 (**96**, Part III, p. 517).

(4) Rice, S. O.: 'Statistical Properties of a Sine Wave Plus Random Noise', *Bell System Technical Journal*, 1948, **27**, p. 109.

(5) C.C.I.R. Report No. 27, London, 1953.

(6) Ratcliffe, J. A.: 'A Survey of Existing Knowledge of Irregularities and Horizontal Movements in the Ionosphere'. Conference on the Physics of the Ionosphere, Cambridge 1954—Physical Society Report, 1955, p. 88.

(7) Briggs, B. H., and Phillips, G. J.: 'Horizontal Irregularities of the Ionosphere', *Proceedings of the Physical Society* B, 1950, **63**, p. 907.

(8) Briggs, B. H., and Spencer, M.: 'Horizontal Movements in the Ionosphere', *Reports on Progress in Physics*, **17**, 245 (The Physical Society, 1954).

(9) Booker, H. G., Ratcliffe, J. A., and Shinn, D. H.: 'Diffraction from an Irregular Screen, with Applications to the Ionosphere Problems', *Philosophical Transactions of the Royal Society* (A), 1950, **262**, p. 579.

(10) Slack, Miss M.: 'The Probability Distributions of Sinusoidal Oscillations Combined in Random Phase', *Journal I.E.E.*, 1946, **93**, Part III, p. 76.

(11) Pearson, K.: 'A Mathematical Theory of Random Migration', Drapers' Company Research Memoirs, Biometric Series III, 1906.

(12) Law, H. B.: 'The Signal/Noise Performance Rating of Receivers for Long-Distance Synchronous Radiotelegraph Systems using Frequency Modulation' (see page 124).

(13) Allnatt, J. W., Jones, E. D. J., and Law, H. B.: 'Frequency Diversity in the Reception of Selectively Fading Binary Frequency-Modulated Signals, with special reference to Long-Distance Radiotelegraphy' (see page 98).

(10) APPENDICES

(10.1) Speed of Fading at Vertical Incidence due to Horizontal Movement of the Ionosphere

Considering the diffraction pattern produced on the ground by the reflection of a normally incident wave in an irregular ionosphere, Briggs and Phillips[7] define the size of the irregularities in terms of a length l, equal to the distance between two points on the ground for which the spatial autocorrelation falls to 0·5, the incident wave being a plane wave. If, instead of being a plane wave, the incident signal is derived from a point source on the ground, the distance on the ground corresponding to a value of 0·5 for the autocorrelation function becomes $2l$. If the reflecting layer drifts horizontally with a velocity u_w the amplitude of a signal received by a stationary aerial will show autocorrelation in time falling to 0·5 in a time

$2l/2u_w$. Assuming that the time autocorrelation function is of Gaussian form, $\rho(t)$ say, given by

$$\rho(t) = \exp(-t^2/2T^2)$$

we have
$$0\cdot 5 = \exp(-l^2/2u_w^2T^2)$$

whence
$$T = 0\cdot 85 l/u_w$$

so that
$$\rho(t) = \exp[-\tfrac{1}{2}t^2(0\cdot 85l/u_w)^{-2}] \quad . \quad . \quad . \quad (4)$$

Booker, Ratcliffe and Shinn[9] show that the time autocorrelation function is approximately equal to the square of the Fourier transform of the function $W(f+f_0)$, where $W(f)$ represents the power spectrum of the received signal and f_0 is the centre frequency of the distribution. Thus we have

$$\text{Fourier transform of } W(f+f_0) = \exp\{\tfrac{1}{2}[-\tfrac{1}{2}t^2(0\cdot 85l/u_w)^{-2}]\}$$

$$W(f+f_0) = A\exp[-\tfrac{1}{2}f^2 2(0\cdot 85l/u_w)^2 4\pi^2]$$

$$W(f) = A\exp[-\tfrac{1}{2}(f-f_0)^2\sigma^{-2}] \quad . \quad . \quad . \quad (5)$$

where
$$\sigma = u_w/2\cdot 4\pi l \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Rice[4] shows that if the power spectrum is symmetrical, as in the present case, the frequency, $N(v)$ say, of upward crossing by the signal envelope of an amplitude v is given by

$$N(v) = (b_2/2\pi)^{1/2}\,(\text{probability density of } v) \quad . \quad . \quad (7)$$

and that if the power spectrum $W(f)$ is of Gaussian form, with standard deviation σ, the constant b_2 is given by

$$b_2 = 4\pi^2\sigma^2V^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where V is the r.m.s. amplitude of the signal.

In the present case the signal conforms to the Rayleigh distribution, so that the probability $\mathrm{P}(v)$ of the signal amplitude being less than v is given by

$$\mathrm{P}(v) = 1 - \exp(-v^2/V^2)$$

and for the probability density we have

$$d\mathrm{P}(v)/dv = \frac{2v}{V^2}\exp(-v^2/V^2) \quad . \quad . \quad . \quad . \quad (9)$$

Substituting in eqns. (7) according to eqns. (8) and (9), we have

$$N(v) = 2(2\pi)^{\frac{1}{2}}\sigma\frac{v}{V}\exp(-v^2/V^2)$$

The maximum value of $N(v)$ occurs at the most probable amplitude $V/2^{1/2}$, so that for this maximum value, N say, which is referred to in Section 2 as the quasi-frequency of fading, we have

$$N = 2\pi^{\frac{1}{2}}\sigma\exp(-\tfrac{1}{2})$$
$$= 2\cdot 15\sigma \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

Thus the relationship between the frequency of upward crossing of an amplitude v and the maximum frequency is

$$N(v)/N = 2^{1/2}\frac{v}{V}\exp\left(\frac{1}{2} - \frac{v^2}{V^2}\right) \quad . \quad . \quad . \quad (11)$$

Inserting in eqn. (10) the value of σ given in eqn. (6), we have, for the fading quasi-frequency at vertical incidence,

$$N = 0\cdot 29u_w/l \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

(10.2) Proportion of Time for which the Signal from One of Two Paths is the Larger

Two signals of instantaneous powers p_1 and p_2 and mean powers P_1 and P_2, fading independently according to the Rayleigh distribution, reach a receiver simultaneously. The probability $\mathrm{P}(p_1)$ that the power of the first signal will be less than p_1 is given by

$$\mathrm{P}(p_1) = 1 - \exp(-p_1/P_1)$$

Similarly,
$$\mathrm{P}(p_2) = 1 - \exp(-p_2/P_2)$$

The probability that the first signal will be less than p_2 while the second signal lies in the interval p_2 to $p_2 + dp_2$ is given by

$$\mathrm{P}(p_1)\Big]_{p_1=p_2} d\mathrm{P}(p_2) = [1 - \exp(-p_2/P_1)]\frac{1}{P_2}\exp(-p_2/P_2)dp_2$$

Integrating this expression over the full range of p_2, from 0 to ∞, we obtain, for the probability that p_1 will be less than p_2 at any time,

$$\mathrm{P}(p_1 < p_2) = \left(1 + \frac{P_1}{P_2}\right)^{-1}$$
$$\simeq P_2/P_1 \text{ if } P_2 \ll P_1$$

Suppose now that p_1 fades as above but that p_2 is fixed in level at the power P_2. The probability that p_1 will be less than p_2 is then given by

$$\mathrm{P}(p_1 < p_2) = 1 - \exp(-P_2/P_1)$$
$$\simeq P_2/P_1 \text{ if } P_2 \ll P_1$$

Thus, provided that the mean power of one signal is very much smaller than that of the other, the probability of the one signal exceeding the other is the same whether the smaller signal is steady or fading.

[The discussion on the above paper will be found on page 147.]

621.396.812 : 621.396.215

Paper No. 2103 R
July 1956

THE SIGNAL/NOISE PERFORMANCE RATING OF RECEIVERS FOR LONG-DISTANCE SYNCHRONOUS RADIOTELEGRAPH SYSTEMS USING FREQUENCY MODULATION

By H. B. LAW, B.Sc.Tech., Associate Member.

(*The paper was first received* 13*th October*, 1955, *and in revised form* 5*th March*, 1956. *It was published in July*, 1956, *and was read before the* RADIO AND TELECOMMUNICATION SECTION 14*th November*, 1956.)

SUMMARY

Experimental results, supported by theory, indicate that the error liability after regeneration of the output of a limiter-discriminator frequency-modulation radiotelegraph receiver, when fed with steady signals plus noise, can be described in terms of a simple exponential relation involving a single parameter, which characterizes the receiver performance. This parameter, defined as the signal/noise energy ratio required to give an error liability of $1/2\varepsilon$, is also the amount by which the receiver falls short of the ideal in the non-diversity detection of Rayleigh-fading signals in noise. It is therefore a convenient index of performance. The losses entailed in diversity by selection with this type of receiver as compared with ideal diversity combination, in the reception of fading signals, are determined for 2-, 3-, and 4-path diversity; they are small, but perhaps not insignificant. Test results suggest that the practice, at present almost universal, of specifying receiver performance in terms of telegraph distortion, with an allowance for associated line tails, causes an appreciable waste of channel capacity.

LIST OF SYMBOLS

a = Characteristic signal/noise power ratio, giving error rate of $1/2\varepsilon$.
b = Characteristic signal/noise energy ratio, giving error rate of $1/2\varepsilon$.
B = Telegraph speed, in bauds.
f = Noise bandwidth.
q = Number of diversity branches.
N = Noise power.
N_0 = Noise power per unit bandwidth.
p = Instantaneous signal power.
P = Mean signal power.
$\mathrm{P}(p)$ = Probability that the signal power is less than p.
P_c = Probability of 'capture' of limiter by noise.
P_e = Probability of error (steady signals).
$\bar{\mathrm{P}}$ = Probability of error (fading signals).
w = Energy in a signal element.
W_0 = Mean signal energy per element.

(1) INTRODUCTION

When complete transmission systems, or items of transmission equipment, are specified or tested the ultimate object is to define the performance in terms having a direct relation with the normal use of the apparatus. Most telegraph systems employ direct printing methods and the practical criterion of performance is the proportion of errors appearing in the printed copy, so that it might be expected that specification would be in terms of error liability. It is, however, more usual to deal in terms of telegraph distortion, for this facilitates the subdivision of tolerances in terminal and transmission equipment. In line practice, a complete telegraph channel may be made up of a number of sections operating in tandem, each section contributing to the overall telegraph distortion; so far as distortion due to fluctuation noise is concerned, the contributions of individual sections add up on an r.m.s. basis and it is clearly appropriate to specify and measure section performance in terms of telegraph distortion. High-frequency radiotelegraphy is superficially similar in that a complete channel consists of a tandem connection of a number of links, commonly three, namely a line link from the sending terminal to the radio transmitter, the radio link itself, and a line link from the radio receiver to the receiving terminal; the custom hitherto has been to follow line practice and to specify the performance of the radio equipment in terms of telegraph distortion. However, the radio case differs fundamentally from the line case, for it is certainly uneconomical, and perhaps even impossible, to satisfy a close distortion tolerance at all times. Radiocommunication is subject to fading, atmospherics, interference and multi-path propagation, all of which cause distortion, and the fact has to be faced that the distortion will sometimes be so great as to cause failure. The overall radiotelegraphy problem is to minimize failures while maintaining proper economy in plant, operation, and bandwidth occupancy.

Since many of the troubles that afflict a radio link are outside the control of the operating organization and are at times so great as to cause errors, it is reasonable to demand that the associated line links, where, in principle, conditions are completely controllable, be of such high quality as to contribute insignificantly to the total error liability. Alternatively, and probably better, the signals might be regenerated at the radio stations, so allowing comparatively wide distortion limits for the line tails. On this argument it is reasonable to test a radiotelegraph system by measuring its own error liability, without any line links. This has sometimes been done in the past, using teleprinters to receive a test message, but the arrangement is open to the objections that it is inflexible in speed of signalling and that the teleprinter margin, i.e. the amount of distortion that it can tolerate, is apt to be uncertain.

It is obviously desirable that any arrangement for overall tests on radiotelegraph equipment should correspond closely with the practical conditions of operation. The great bulk of the point-to-point radiotelegraph channels handled by Post Office stations in the United Kingdom are of the synchronous kind, operating in the speed range 80–100 bauds, each radio channel giving two 40–50-baud operator-to-operator channels by time-division multiplex. The distributors used to separate the channels operate on a sampling basis, with inspection times of the order of 1–2 millisec in the case of mechanical distributors, which at present greatly predominate; electronic distributors having sampling periods of the order of 0·1 millisec are also used. Thus the time-division multiplexing implies regeneration at the receiving terminal and this leads to the idea of testing radiotelegraph systems by counting the errors in the received regenerated signals. This method has the advantages that results are readily interpretable and that questions of distortion margin do not arise, and it lends itself to simple automatic error-counting techniques. Furthermore, it seems safe to base the testing technique on regeneration because its advantages are so great that it is sure to be used more and more

Mr. Law is at the Post Office Research Station.

for its own sake, quite apart from its presence as a by-product in time-division multiplex systems. In fact, it can be argued that the process of reception of a radiotelegraph signal is incomplete until the signal has been regenerated, and that the proper place to do the regeneration is at the radio receiving station, where the maximum of information about the signal is available.

The considerations just outlined have led to the development of equipment for determining the performance of radiotelegraph systems in the laboratory by comparing the received regenerated signals element-by-element with those actually sent, and counting the elements in error. The test equipment caters for the injection of white noise, and the signal level can either be held steady or be caused to vary according to the Rayleigh distribution characteristic of the fading of long-distance high-frequency signals; diversity reception is catered for. The equipment has been in use for some months, on several different radiotelegraph systems, and the experience so gained, coupled with theoretical work, has led to a method of describing noisy-signal system performance that is succinct and also absolute, in the sense of being based on the best standard of performance theoretically attainable.

The present paper is concerned with the application of this method of describing performance to binary radiotelegraph systems of the frequency-modulated type, in which information is conveyed by shifting the transmitter frequency to and fro between two values corresponding to the mark and space conditions and differing usually by a few hundred cycles per second. The problem is mainly one concerning the later stages of the radiotelegraph receiver, for up to a certain point in the final intermediate-frequency stage the conventional receiver behaves in a substantially linear fashion and its performance can be described in well-understood terms of noise factor, selectivity and so forth. Beyond this point the signals are demodulated and converted into square d.c. telegraph signals of constant amplitude, the overall process being essentially non-linear. Apart from the dependence of telegraph distortion upon signal/noise ratio, for reasonably good values of the ratio which give the full 'frequency-modulation improvement', the noisy-signal performance of radiotelegraph receivers has been inadequately understood and has been described in such crude terms as 'the signal strength required to give a clean mark'. It has already been suggested that distortion criteria are inappropriate for long-distance high-frequency radiotelegraphy; also, with fading signals the performance in the range of signal/noise ratios below that which can be described as reasonably good is of high practical importance. Application of the new method, which is free from such restrictions, is discussed in the following Sections first for the steady-signal case and then for the fading-signal case, with and without diversity.

(2) STEADY-SIGNAL PERFORMANCE

The later stages of a practical frequency-modulation radiotelegraph receiver according to current ideas consist essentially of an intermediate-frequency channel filter, a limiter to get rid of amplitude fluctuations, and some kind of frequency discriminator followed by a low-pass filter and squaring stages to produce d.c. telegraph signals. In general, some of these items will be duplicated to provide for diversity reception, with switching or combination at some point, but that is an aspect of no concern at this stage of the discussion. Considering now the noisy-signal performance of such a receiver, it is evident that the limiter loses the signal, or is 'captured' by noise, whenever a noise-envelope peak exceeds the signal amplitude. The probability of this occurring, P_c, is given by the Rayleigh formula

$$P_c = \exp(-p/N)$$

where p is the signal power and N is the noise power. Assuming that the output-voltage/input-frequency characteristic of the discriminator is antisymmetrical about a balance point midway between the mark and space frequencies, and that the loss/frequency curve of the channel filter is symmetrical about the same point, there is an even chance that any noise peak will change the sign of the discriminator output. The stated assumptions are substantially realized in normal practice. Thus if the telegraph output were obtained by single-point sampling at the discriminator output the overall probability of error, P_e, would be

$$P_e = \tfrac{1}{2}\exp(-p/N)$$

This is the result obtained by Montgomery.[1] In the case in question the post-discriminator filter reduces the probability of error, and as a first guess the filter might be expected to give a straightforward improvement in signal/noise ratio, analogous to its effect when the input signal sufficiently exceeds the noise. Thus the expression for error liability might take the form

$$P_e = \tfrac{1}{2}\exp(-p/aN) \quad . \quad . \quad . \quad . \quad . \quad (1)$$

wherein the factor a describes the improvement effected by the post-discriminator filter.

The graphs given in Figs. 1, 2 and 3 show the results of steady-

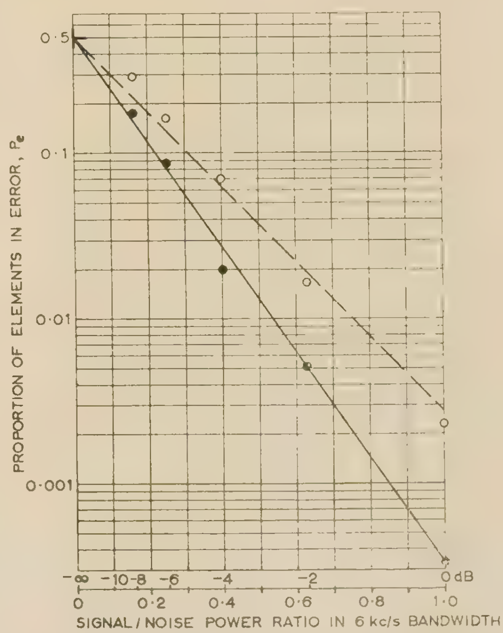


Fig. 1.—Steady-signal performance—System A.

——— 50 bauds; $a = -8{\cdot}6$ dB; $b = 12{\cdot}2$ dB.
— — — 110 bauds; $a = -7{\cdot}1$ dB; $b = 10{\cdot}3$ dB.

signal tests on three different frequency-modulation telegraph receivers. The signal/noise power-ratio scale and the proportion-of-error scale are respectively linear and logarithmic, although, for practical convenience in plotting attenuator settings, the former is normally calibrated in decibels, as shown at the bottom of the graphs. It will be seen that for each system and set of conditions the measured points lie along a straight line passing through the point $(0, 0{\cdot}5)$. Thus there is justification in using eqn. (1) to describe the performance of the receivers. The values of a are given in the graphs. The measurements of signal/noise power ratio made in obtaining the results were invariably in a bandwidth of 6 kc/s, and the quoted power ratios always refer to this bandwidth; the actual bandwidth at the limiter input was 2 kc/s for the results given in Figs. 1 and 2, and 1 kc/s for those in Fig. 3.

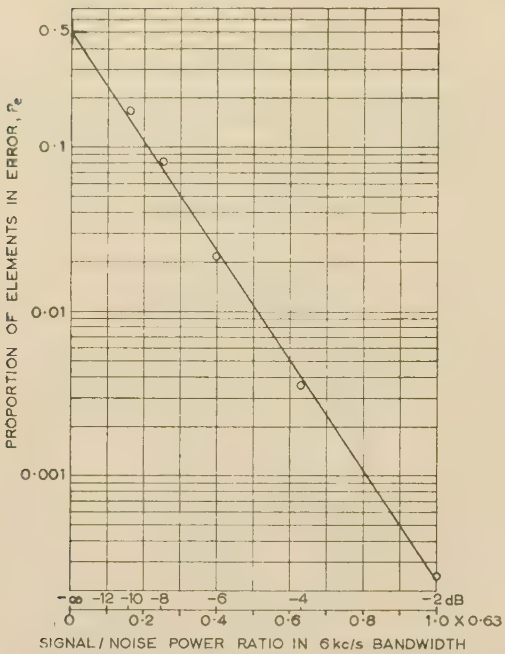


Fig. 2.—Steady-signal performance—System B.
100 bauds; $a = -10{\cdot}9$ dB; $b = 6{\cdot}9$ dB.

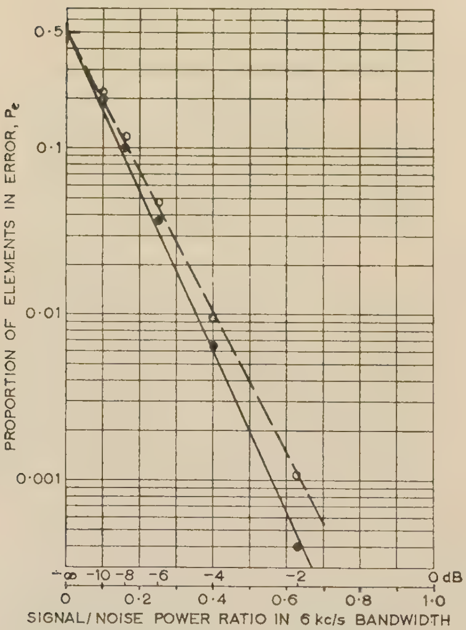


Fig. 3.—Steady-signal performance—System C.
——— 100 bauds; $a = -10{\cdot}3$ dB; $b = 7{\cdot}5$ dB.
– – – 200 bauds; $a = -9{\cdot}7$ dB; $b = 5{\cdot}1$ dB.

Many results giving straight lines similar to those in Figs. 1, 2 and 3 have been obtained. Curvilinear characteristics have also been obtained on occasion, but have always been associated with some defect in design or adjustment which has affected the symmetry of the receiver characteristics relative to the balance point of the discriminator. Thus there appears to be justification for using eqn. (1) to describe receiver performance, which may therefore be characterized by the single quantity a for any particular set of conditions. This quantity may be referred to as the characteristic signal/noise power ratio, and defined as the power ratio that gives an error liability of $1/2\varepsilon$.

A more fundamental quantity than the characteristic signal/noise power ratio is the corresponding ratio of the energy in a signal element to the noise power unit bandwidth; this is referred to as the characteristic signal/noise energy ratio, b, given by

$$b = af/B$$

where f is the noise bandwidth in cycles per second in which the signal/noise power ratio is determined, and B is the telegraph speeds in bauds. Thus eqn. (1) may be rewritten in the form

$$P_e = \tfrac{1}{2}\exp(-w/bN_0) \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where w is the energy in a signal element, and N_0 is the noise power per unit bandwidth, which also has the dimensions of energy. The characteristic signal/noise energy ratio may now be defined as the ratio of signal energy per element to noise power per unit bandwidth that gives an error liability of $1/2\varepsilon$.

Following the methods of Woodward[2] and Davies[2,3] it can be shown[4] that the minimum error liability theoretically attainable in a two-tone or frequency-modulation system is given by

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\,(w/N_0)^{\frac{1}{2}} \quad . \quad . \quad . \quad . \quad (3)$$

provided that the signal energy per element is the same for mark and space and that mark and space are equally likely to be sent; practical messages in the codes commonly used satisfy the latter proviso approximately. This theoretical limit to performance provides a convenient standard in terms of which a practical system may be assessed. The amount by which the performance of an equipment falls short of the ideal may be expressed as the ratio of the signal energy required to produce a given error rate in the practical equipment to that required for the same error rate in the ideal receiver, the noise being the same in the two cases. The term 'demodulation factor' is suggested elsewhere[4] for this measure of receiver imperfection.

The curve in Fig. 4 shows the ideal performance plotted against the linear signal/noise and logarithmic error-liability

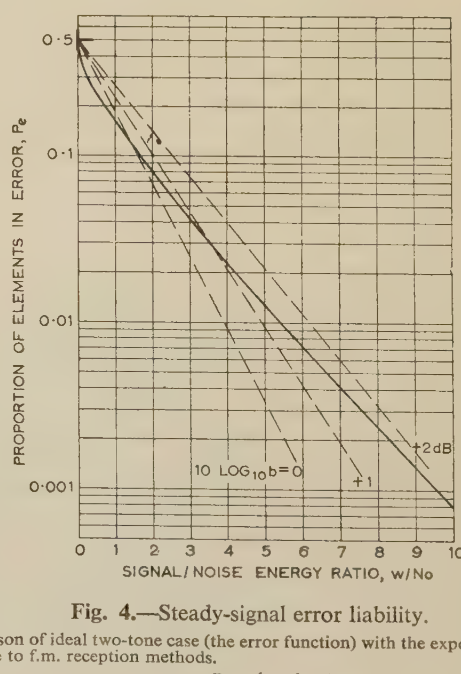


Fig. 4.—Steady-signal error liability.
Comparison of ideal two-tone case (the error function) with the exponential function appropriate to f.m. reception methods.
——— $P_e = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\,\sqrt{(w/N_0)}$.
– – – $P_e = \frac{1}{2}\exp(-w/bN_0)$.

scales that have been found suitable for the practical frequency-modulation case. It would be convenient if the ideal curve could be replaced by an equivalent straight line passing through the point (0, 0·5), but this could only be an approximation; in other words, the demodulation factor of a receiver having an exponential relation between error rate and signal/noise ratio varies appreciably with error rate. This difficulty disappears when the case of fading signals is considered.

It is of interest to consider how the characteristic signal/noise power and energy ratios vary with the speed of signalling. The graphs in Fig. 5 show some typical results. It will be seen that

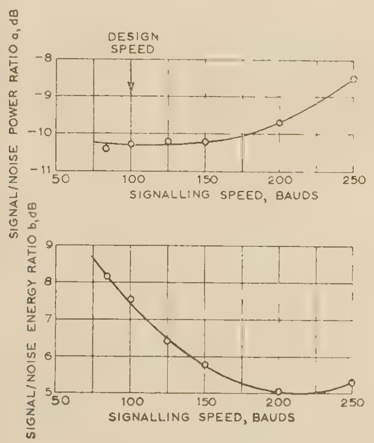


Fig. 5.—Effect of signalling speed on performance—System C.

The curves show the characteristic signal/noise power ratio a (in 6 kc/s bandwidth) and energy ratio b as functions of the speed of signalling.

as the signalling speed is raised from a low value the power ratio remains substantially constant until a point is reached, determined by the post-discriminator filter characteristic, where appreciable amounts of signal power become lost and the performance deteriorates. So long as the power ratio remains constant the energy ratio improves proportionately with increasing speed and then, soon after the post-discriminator filter begins to affect the signal, an optimum is reached. It may be remarked in passing that the optimum speed determined in this way is about twice the design value based on distortion for the system in question; similar results have been obtained in another case in which a comparison of this kind was made. This gives some idea of the waste of channel capacity involved in applying line specification methods to radio.

(3) FADING-SIGNAL PERFORMANCE

(3.1) Theoretical Considerations

Long-distance radio signals in the 4–30 Mc/s band are subject to fading, and it appears that for periods of the order of five minutes the amplitude distribution is adequately described by the Rayleigh formula. Thus the proportion of time P(p) for which the received signal power is less than p is given by

$$\mathrm{P}(p) = 1 - \exp(-p/P) \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where P is the mean signal power. Hence the proportion of time during which the signal power lies between the limits p and $p + dp$ is given by

$$d\mathrm{P}(p) = \frac{1}{P}\exp(-p/P)dp$$

If now we assume that the signal strength can be regarded as constant during any one signal element, eqn. (1) can be used to give the probability density of errors at the signal level p, thus

$$\mathrm{P}_e d\mathrm{P}(p) = \tfrac{1}{2}\exp(-p/aN)\frac{1}{P}\exp(-p/P)dp$$

and the overall probability of error, P, is obtained by integrating this expression:

$$\mathrm{P} = \int_0^\infty \frac{1}{2P}\exp(-p/P - p/aN)dp$$

$$= \frac{\tfrac{1}{2}}{1 + \dfrac{P}{aN}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

or alternatively, in terms of the mean signal/noise energy ratio,

$$\mathrm{P} = \frac{\tfrac{1}{2}}{1 + W_0/bN_0} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

where W_0 is the mean signal energy per element.

The performance of an ideal on/off system on Rayleigh-fading signals, without diversity, has been calculated.[4] Assuming that the fading is not frequency-selective ('flat fading'), that mark and space are equally likely to be sent and that the *mean* signal energy per element and noise energy are the same in the two cases, the performances of an ideal two-tone or frequency-modulation system and an ideal on/off system will be identical. Three points for the ideal case have been plotted in Fig. 6. It will be

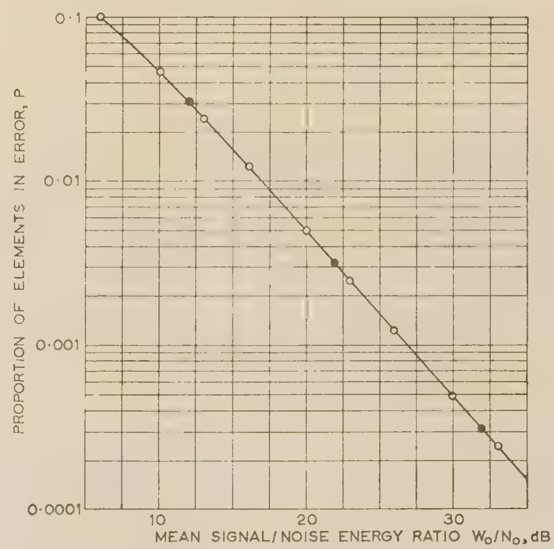


Fig. 6.—Ideal performance with flat-fading signal—no diversity.

● Values of P for ideal performance characteristic.

○ Values of P from the expression $P = \dfrac{\tfrac{1}{2}}{1 + W_0/N_0}$.

seen that they lie precisely on the curve corresponding to eqn. (6) for the case $b = 1$. Thus it turns out that the characteristic signal/noise energy ratio, b, which, it will be recalled, was defined as the signal/noise energy ratio required to give an error rate of $1/2\varepsilon$ on steady signals, is exactly the amount by which the system falls short of the ideal in the non-diversity reception of flat-fading signals. The characteristic signal/noise energy ratio is thus equal to the demodulation factor in this case.

Spaced-aerial diversity is normally used in the reception of long-distance point-to-point radiotelegraph signals, the diversity

path which carries the largest signal having control of the receiver output at any instant. Assuming that each diversity path has the same mean signal power P, and that the fadings in the different paths are uncorrelated, the probability that the largest signal in q diversity paths is less than p is

$$\mathrm{P}(p) = [1 - \exp(-p/P)]^q$$

hence
$$d\mathrm{P}(p) = q[1 - \exp(-p/P)]^{q-1}\exp(-p/P)\frac{1}{P}dp$$

Assuming that the noise power in each diversity path is the same and equal to N, and applying eqn. (1), we have for the probability density of error at the signal level p

$$d\mathrm{P} = \tfrac{1}{2}\exp(-p/aN)q[1 - \exp(-p/P)]^{q-1}\exp(-p/P)\frac{1}{P}dp$$

Expansion of the binomial term followed by integration of the expression over the range of p from zero to infinity gives the total error probability P as follows:

$$\mathrm{P} = \frac{q}{2}\left[\frac{1}{1 + P/aN} - \frac{q-1}{1}\frac{1}{2 + P/aN} + \frac{(q-1)(q-2)}{2!}\frac{1}{3 + P/aN}\ldots + (1-)^r\frac{(q-1)(q-2)\ldots(q-r)}{r!}\frac{1}{r+1+P/aN}\ldots + (-1)^{q-1}\frac{(q-1)!}{(q-1)!}\frac{1}{q + P/aN}\right]$$

whence by induction

$$\mathrm{P} = \frac{1}{2}\frac{q!}{(1 + P/aN)(2 + P/aN)\ldots(q + P/aN)} \quad . \quad (7)$$

or in terms of energy ratio

$$\mathrm{P} = \frac{1}{2}\frac{q!}{(1 + W_0/bN_0)(2 + W_0/bN_0)\ldots(q + W_0/bN_0)} \quad . \quad (8)$$

Values of P have been plotted according to this expression for $q = 1, 2, 3$ and 4 in Fig. 7. Again the curves are very like those of the ideal case,[4] but the diversity gains are a little less than ideal; Fig. 8 shows the diversity gains as functions of the grade of service for the two cases. It will be seen that for reasonably low error liabilities the selection-diversity gain in dual diversity is about 1 dB less than ideal. This is a measure of the loss entailed in selection as compared with ideal combination,* and it is interesting to compare this figure with the corresponding figure of 3 dB for the steady-signal case.[5]

(3.2) Typical Results

Fig. 9 gives the results of measurements under laboratory-simulated fading conditions of the system having the steady-signal performance given in Fig. 2. The curves were obtained by calculation from eqn. (8), using the value of characteristic signal/noise energy ratio derived from the steady-signal performance, namely $b = 6{\cdot}9$ dB. It will be seen that the fit of the measured points to the calculated curve is, in general, good. This shows incidentally that the diversity gain possible with this type of system is fully realized. The point (25·6, 0·000 37) which lies some way from the curve represents the average of error counts of 6, 1 and 4 in three successive runs of 10 000 elements; counts as small as these are subject to considerable uncertainty and a reliable result is obtainable only by greatly prolonging the test.

The results obtained under the severe selective fading condition given by two-path propagation with a path time-delay difference of 2 millisec, the two paths being equally active, show very little difference from the flat-fading case. The dual-diversity results show a slight deterioration and this change is in the expected direction. The apparent slight improvement in the non-diversity case is probably spurious and attributable to inaccuracies of measurement. The possibility of obtaining a frequency-diversity advantage under selective fading conditions is discussed in another paper.[6]

* See, however, the Appendix to Paper No. 2104 (page 140.)

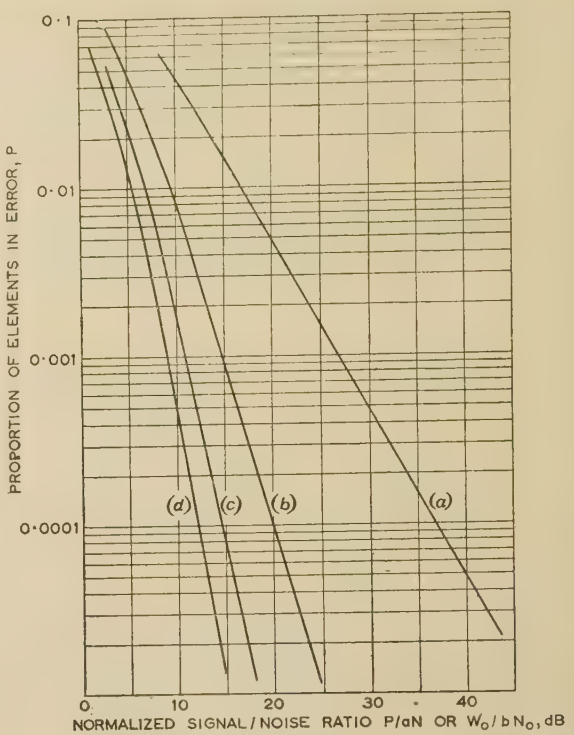


Fig. 7.—Fading-signal performance for diversity operation on a selection basis.

(*a*) No diversity.
(*b*) 2 diversity branches.
(*c*) 3 diversity branches.
(*d*) 4 diversity branches.

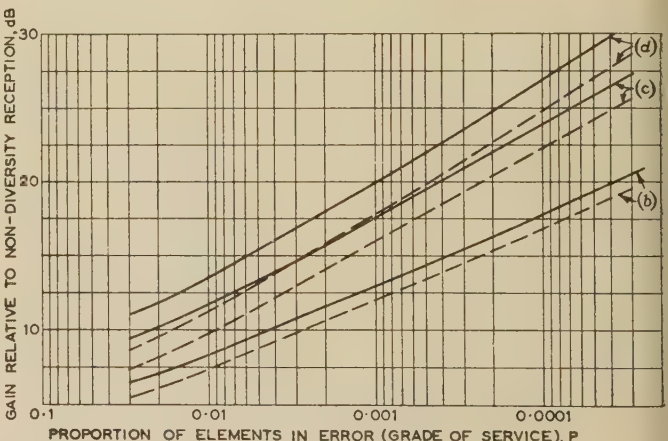


Fig. 8.—Diversity gains.

——— Ideal case.
— — — Branch selection on a signal-amplitude basis.
(*b*) 2 diversity branches.
(*c*) 3 diversity branches.
(*d*) 4 diversity branches.

(4) DISCUSSION AND CONCLUSIONS

It appears from the work done so far that the steady-signal error-liability of frequency-modulation radiotelegraph receivers

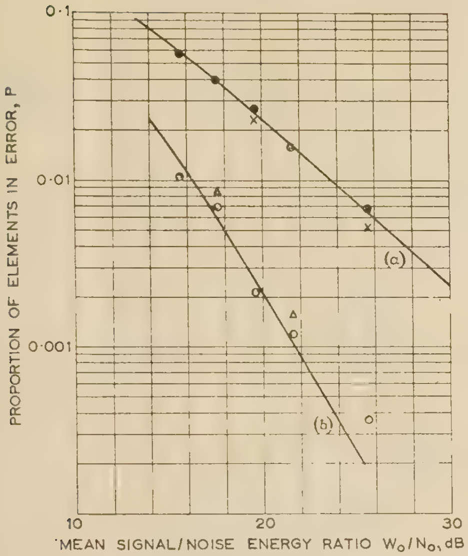


Fig. 9.—Results of tests with fading signals—System B, 100 bauds.

Curves are calculated, with an assumed characteristic signal/noise energy ratio = 6·9 dB.
(*a*) No diversity.
(*b*) Dual diversity.

Points are measured for following conditions:
● Flat fading, no diversity.
× Selective fading (path time delay = 2 millisec) no diversity.
○ Flat fading, dual diversity.
△ Selective fading (path time delay = 2 millisec) dual diversity.

of the limiter-discriminator type can be described in terms of a simple exponential function of the signal/noise ratio and that the fading-signal performance, with or without diversity, is readily derived from the steady-signal performance. Although the exponential relation has been derived empirically from test results on three systems only, there is a fair measure of theoretical support for it, and it seems reasonable to assume that it is generally valid. Obviously, it should be checked as a routine matter whenever a new frequency-modulation system is subjected to laboratory tests.

On the assumption that the relation does in fact hold, the performance of a receiver in detecting a steady signal in the presence of noise can be completely described, for a given signalling speed, in terms of its characteristic signal/noise energy ratio, which is defined as the ratio of the signal energy per telegraph element to the noise power per unit bandwidth required to give an error rate of $1/2\varepsilon$. This ratio is of particular significance, for it is also the amount by which the receiver falls short of the ideal in detecting a flat-fading signal in non-diversity operation. The effectiveness of any diversity switching arrangement can be assessed by comparison with the diversity-by-selection curves and formulae that have been derived. The interesting point emerges in this connection that, with receivers of the limiter-discriminator type, dual diversity with branch selection on an amplitude basis, if perfect of its kind, gives about 1 dB less gain than the ideal (combination) diversity; the corresponding disadvantage in a comparison of the quadruple-diversity cases is about 2 dB. The advantage to be gained by adopting combination methods is evidently small, although perhaps not insignificant.

The optimum working speeds obtained with systems tested on the basis of regenerated-signal error liability tend to be higher, by a factor of the order of two, than the design maximum speeds based on distortion considerations. This indicates the important loss of channel capacity that can arise from the application to radiotelegraphy of line-telegraphy methods of specification, probably including stringent limitations on characteristic distortion. The root of the difficulty lies in the necessity for distortion allowances for the line tails associated with the radio link; yet it seems unreasonable that the performance of perhaps many thousands of miles of radio route should be prejudiced by, at most a few hundred miles of line. Furthermore, the expenditure involved in eliminating that necessity, e.g. by the provision of regenerative telegraph repeaters at the radio stations, would be small in relation to the capital already invested in radio transmitters, receivers, aerials, line equipment and terminal equipment. There is here a cheap way of increasing channel capacity.

(5) ACKNOWLEDGMENTS

The author's thanks are due to Messrs. F. J. Lee and F. A. W. Levett for the experimental results.

Acknowledgment is made to the Engineer-in-Chief of the Post Office and to the Controller of H.M. Stationery Office for permission to publish the paper.

(6) REFERENCES

(1) MONTGOMERY, G. F.: 'Comparison of Amplitude and Phase Modulation for Narrow-Band Communication of Binary Coded Messages in Fluctuation Noise', *Proceedings of the Institute of Radio Engineers*, February, 1954, **42**, p. 447.
(2) WOODWARD, P. M., and DAVIES, I. L.: 'Information Theory and Inverse Probability in Telecommunication', *Proceedings I.E.E.*, Paper No. 1239 R, March, 1952 (**99**, Part III, p. 37).
(3) DAVIES, I. L.: 'On Determining the Presence of Signals in Noise', *ibid.*, Paper No. 1256 R, March, 1952 (**99**, Part III, p. 45).
(4) LAW, H. B.: 'The Detectability of Fading Radiotelegraph Signals in Noise' (see next page).
(5) KAHN, L. R.: 'Ratio Squarer', *Proceedings of the Institute of Radio Engineers*, November, 1954, **42**, p. 1704.
(6) ALLNATT, J. W., JONES, E. D. J., and LAW, H. B.: 'Frequency Diversity in the Reception of Selectively Fading Binary Frequency-Modulated Signals with special reference to Long-Distance Radiotelegraphy' (see page 98).

[The discussion on the above paper will be found on page 147.]

621.396.812 : 621.396.215

Paper No. 2104 R
July 1956

THE DETECTABILITY OF FADING RADIOTELEGRAPH SIGNALS IN NOISE

By H. B. LAW, B.Sc.Tech., Associate Member.

(*The paper was first received* 13*th October*, 1955, *and in revised form* 5*th March*, 1956. *It was published in July*, 1956, *and was read before the* Radio and Telecommunication Section 14*th November*, 1956.)

SUMMARY

An ideal receiver for binary synchronous telegraphy is postulated; this receiver is defined as one that interprets each element of a received signal with the minimum probability of error. The relations between error liability and signal/noise ratio are determined for steady signals and for Rayleigh-fading signals in white Gaussian noise, using up to four diversity branches. The noisy-signal performance of a practical receiver can be described in terms of the amount by which it falls short of the ideal, and it is proposed that this measure of imperfection, expressed as an energy ratio, be known as the 'demodulation factor'.

The analysis leads to a mathematical specification for the ideal diversity receiver, and this provides a starting-point for the design of practical receivers. The outputs of diversity branches are combined, weighted according to signal energy, instead of the largest output being selected. Very deep fading proves to be of little importance. It is found that the use of a 7-unit error-detecting code in place of an unprotected 5-unit code is roughly equivalent to doubling the number of diversity branches. Cross-correlation between the mark and space signals in two-tone or frequency-shift systems is found to be insignificant under the conditions ordinarily encountered in long-distance radiotelegraphy.

Although it is primarily concerned with high-frequency radiotelegraphy, the study may prove useful in other fields, including that of microwave pulse communication.

LIST OF SYMBOLS

f_c = Centre frequency of a two-tone or frequency-shift signal.
f_d = Frequency difference between mark and space.
m_u = Undistorted mark-signal waveform in the uth mark branch.
n = Noise waveform.
N_0 = Noise power per unit bandwidth.
$\mathrm{p}(\varepsilon|y)$ = Probability of the existence of a signal, y having been received.
$\mathrm{p}(\varepsilon_m|y)$ = Probability of the existence of mark, y having been received.
$\mathrm{p}(w)$ = Probability that the signal-on energy in a branch is less than w.
$\mathrm{p}(y|s)$ = Probability of receiving y if a signal is sent.
$\mathrm{p}(y|0)$ = Probability of receiving y if no signal is sent.
$\mathrm{p}(y_{mu}|M)$ = Probability of receiving y_{mu} if mark is sent.
$\mathrm{p}(y_{mu}|S)$ = Probability of receiving y_{mu} if space is sent.
$\mathrm{p}(y_{sv}|M)$ = Probability of receiving y_{sv} if mark is sent.
$\mathrm{p}(y_{sv}|S)$ = Probability of receiving y_{sv} if space is sent.
$\mathrm{p}(\Sigma y|M)$ = Probability of receiving all the waveforms y together if mark is sent.
$\mathrm{p}(\Sigma y|S)$ = Probability of receiving all the waveforms y together if space is sent.
P_e = Element error probability—steady signals.
P_{eq} = Element error probability—reception of fading signals in q diversity branches.
P_{cdr} = Probability of a detected character error.
P_{cur} = Probability of an undetected character error.
$\mathrm{P}_{c0r}, \mathrm{P}_{c1r}$ = Probability of receiving a character of r elements with 0, 1 . . . elements in error.
p, q = Number of diversity branches.
s = Undistorted signal waveform.
s_v = Undistorted space-signal waveform in the vth space branch.
T = Duration of signal element.
w = Energy in a signal element.
w_e = Effective signal energy per element.
W_{co} = Effective character energy per branch averaged over many fading periods.
W_0 = Effective signal-element energy per branch averaged over many fading periods.
w_{mu}, w_{sv} = Signal element energies in the uth mark and vth space branches.
y = Received waveform, noise plus signal (if any).
y_{mu}, y_{sv} = Received waveforms in the uth mark and vth space branches.
λ = *A priori* probability of mark being received.
μ = *A priori* probability of space being received.
ρ = Cross-correlation between mark and space signals.

Mr. Law is at the Post Office Research Station.

(1) INTRODUCTION

Fading and noise are two of the more serious troubles that afflict long-distance radiotelegraphy in the 4–30 Mc/s band, and in point-to-point working their effects are usually mitigated by the use both of spaced-aerial diversity and of receiving equipment of great selectivity. The specification of the performance of such equipment, and its assessment on test, have hitherto usually been in terms of telegraph distortion, but, with the increasing use of synchronous regenerative systems, there is much to be said for using error-liability as the criterion. It is therefore desirable to establish a measure of performance in such terms. Any arbitrary but properly defined standard of performance could be used as a basis for the assessment of practical equipments, but a standard having some absolute significance is to be preferred. A good one for this purpose would be the performance of the best possible system, for measurements made in terms of it would indicate by how much an equipment on test fell short of that ideal, and hence would serve as a valuable guide in development work.

Telegraphy is, of course, a classic case of coded communication, and communication theory provides a means for the calculation of the performance of the ideal receiver fed with signals plus noise. The study recorded in the paper is an application of the theory to the case of the diversity reception of fading two-condition synchronous telegraph signals in the presence of white Gaussian noise.

(2) BASIS OF ANALYSIS

A binary-code telegraph signal consists of a succession of signal elements, each of which must be of one of two kinds, e.g. mark or space. It is assumed in the paper that the essential function of a radiotelegraph receiver is to determine for each element whether it is a mark or a space, with the least probability of error. In order that the risk of error may be minimized the

eceiver must take proper account of all that is known about the ignals, such as the forms of the perfect mark and space elements, he instants of transition between them, the probabilities of their eing transmitted and their amplitudes and carrier phases, as vell as the form of the noise-distorted signal actually received. The analysis adopted is based on the methods used by Woodward[1] and Davies,[1,2] who show that the probability that a received vaveform y was caused by a signal having been sent, or the existence probability' of a signal $p(\varepsilon|y)$, is given by

$$p(\varepsilon|y) = \frac{\lambda p(y|s)}{\lambda p(y|s) + \mu p(y|0)} \quad . \quad . \quad . \quad . \quad (1)$$

where λ = The *a priori* probability of the signal being received.
μ = The *a priori* probability of the signal not being received.
$p(y|s)$ = The probability of the waveform y occurring in the presence of a signal.
$p(y|0)$ = The probability of the waveform y occurring in the absence of a signal.

If the noise is Gaussian and it has a uniform power spectrum over a bandwidth wider than that of the signal, the probability of the waveform y occurring in the presence of a signal, $p(y|s)$, is given by

$$p(y|s) = k \exp\left[-\frac{1}{N_0}\int_0^T (y-s)^2 dt\right] \quad . \quad . \quad . \quad (2)$$

where k is a constant, N_0 is the noise power per unit bandwidth, s is the waveform (assumed to be completely known) that would be produced by the signal alone, i.e. without noise, and T is the duration of the signal. Eqn. (2) shows, to quote Woodward and Davies, 'that, apart from the *a priori* weighting factor, the most probable message is the one whose waveform has the least r.m.s. departure from the received waveform, a result which is certainly intuitive'. Similarly, $p(y|0)$ is given by

$$p(y|0) = k \exp\left[-\frac{1}{N_0}\int_0^T y^2 dt\right] \quad . \quad . \quad . \quad (3)$$

Thus the existence probability may now be reduced to the form

$$p(\varepsilon|y) = \frac{\lambda}{\lambda + \mu \exp\left[\frac{1}{N_0}\int_0^T (s^2 - 2ys)dt\right]}$$

$$= \frac{\lambda}{\lambda + \mu \exp\left[\frac{w}{N_0} - \frac{2}{N_0}\int_0^T ysdt\right]} \quad . \quad . \quad . \quad (4)$$

where w is the energy of the signal, given by

$$w = \int_0^T s^2 dt$$

It should be noted that considerations of circuit impedance are not relevant to the discussion; instead of introducing the impedance only to see it eliminated later, a normalized impedance $(1 + j0)$ has been assumed for the sake of simplicity.

Suppose now that the signal is in fact absent, so that y is composed wholly of noise. In these circumstances the function

$$\frac{2}{N_0}\int_0^T ysdt = \frac{2}{N_0}\int_0^T nsdt = z \text{ (say)}$$

wherein n is the noise waveform, has a Gaussian probability distribution[2] of mean value zero and mean-square value $2w/N_0$:

$$\left.\begin{aligned} \bar{z} &= 0 \\ \overline{z^2} &= 2w/N_0 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

From this point Davies goes on to evaluate the mean existence and non-existence probabilities as functions of w/N_0.

An ideal telegraph receiver according to the definition given above would determine for each element, with the minimum probability of error, whether it was a mark or a space, and the present problem is to find how this minimum probability of error varies with signal/noise ratio. Clearly, the best that the receiver can do is to interpret a signal element as mark or space according as the existence probability of mark is greater or less than 0·5. In plain-language messages the *a priori* probability of each element is affected by what has gone before, according to the frequencies of different letters, words and sequences of words in the language. The ideal message receiver would take account of this information, which, however, is considered to be beyond the scope of the ideal telegraph receiver, and in the present study each element is treated independently. The *a priori* probability of a mark is therefore the same for all elements and is equal to the proportion of marks to total elements in a message. This is about 50% in codes in common use.

Taking as a simple example the reception of on/off signals under no-fading no-diversity conditions, and assuming that mark and space are equally likely to be sent, so that

$$\lambda = \mu = 0{\cdot}5$$

we see by eqn. (4) that the existence probability of mark will exceed 0·5 if

$$\frac{w}{N_0} - \frac{2}{N_0}\int_0^T ysdt < 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

and, using eqns. (5) and the properties of the error function, the probability of this inequality being satisfied if the signal is in fact absent, i.e. the probability of error, P_e, is given by

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\operatorname{erf}\frac{w}{N_0}\left(\frac{N_0}{2w}\right)^{1/2}$$

$$= \tfrac{1}{2} - \tfrac{1}{2}\operatorname{erf}(w/2N_0)^{1/2} \quad . \quad . \quad . \quad . \quad (7)$$

Similarly, if the signal is in fact present,

$$\frac{w}{N_0} - \frac{2}{N_0}\int_0^T ysdt = \frac{w}{N_0} - \frac{2}{N_0}\int_0^T (s^2 + ns)dt$$

$$= -\frac{w}{N_0} - \frac{2}{N_0}\int_0^T nsdt$$

and the probability of error, which is now the probability of expression (6) failing to be satisfied, is the same as before.

In the discussion so far, the signal energy w is the received signal energy when the signal exists. In the present example the signal is a mark telegraph element; since the system is on-off telegraphy the energy in a space element is zero, and since mark and space have been assumed to be equally likely to be sent the average received signal energy per element, $\bar{w}$, is given by

$$\bar{w} = \tfrac{1}{2}w \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

and the expression for the probability of error may be written

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\operatorname{erf}(\bar{w}/N_0)^{1/2} \quad . \quad . \quad . \quad . \quad (9)$$

In long-distance high-frequency radio channels the signals are subject to fading and the signal/noise energy ratio therefore varies with time; fortunately the fading is usually very slow relative to the speed of signalling so that it is reasonable to make the simplifying assumption that the signal amplitude remains constant during any signal element. Also, spaced-aerial diversity is often employed, and this raises the problem of the optimum use of the outputs of the several diversity branches. Finally, two-tone or frequency-shift methods of signalling are commonly used, mark signals being sent on one carrier frequency and space signals on another. The methods of analysis used in the above example are extended to cover the long-distance case in the next Section. For the sake of simplicity the assumption is made that the *a priori* probabilities of mark and space are equal; this assumption is only approximately justified in practice, and Section 3.4 includes a brief discussion of the errors involved in applying it to the important 4/3 codes.

(3) ELEMENT ERROR LIABILITY

(3.1) Generalized Ideal Diversity Combination

Suppose that there are q diversity branches, of which p are for the detection of mark signals in a two-tone system, and $q-p$ are for the detection of space signals; the on-off case is included by making $q = p$. Let the noise power per unit bandwidth N_0 be the same in each branch. Let the form of the undistorted signals in each branch be precisely known, m_u for the uth mark branch and s_v for the vth space branch. Let the received signals in these branches be y_{mu} and y_{sv} respectively. Then by eqn. (2) the probability of receiving y_{mu} when a mark is sent is given by

$$p(y_{mu}|M) = k \exp\left[-\frac{1}{N_0}\int_0^T (y_{mu} - m_u)^2 dt\right]$$

where T is the duration of one signal element and k is a constant as before. The probability of receiving y_{sv} when mark is sent is, by eqn. (3),

$$p(y_{sv}|M) = k \exp\left[-\frac{1}{N_0}\int_0^T (y_{sv})^2 dt\right]$$

Similarly, when space is sent,

$$p(y_{mu}|S) = k \exp\left[-\frac{1}{N_0}\int_0^T (y_{mu})^2 dt\right]$$

$$p(y_{sv}|S) = k \exp\left[-\frac{1}{N_0}\int_0^T (y_{sv} - s_v)^2 dt\right]$$

Hence the probability, $p(\Sigma y|M)$, of receiving $y_{m1}, y_{m2}, \ldots, y_{mp}$ and $y_{s1}, y_{s2}, \ldots, y_{s(q-p)}$ all together, if mark is sent, is given by

$$p(\Sigma y|M) = k^q \exp\left\{-\frac{1}{N_0}\int_0^T\left[\sum_{u=1}^{p}(y_{mu} - m_u)^2 + \sum_{v=1}^{q-p}(y_{sv})^2\right]dt\right\}$$

Similarly, if space is sent

$$p(\Sigma y|S) = k^q \exp\left\{-\frac{1}{N_0}\int_0^T\left[\sum_{u=1}^{p}(y_{mu})^2 + \sum_{v=1}^{q-p}(y_{sv} - s_v)^2\right]dt\right\}$$

On the assumption that mark and space are equally likely to be sent, the existence probability of mark, $p(\varepsilon_m|y)$, is, by eqn. (1),

$$p(\varepsilon_m|y) = \frac{1}{1 + p(\Sigma y|S)/p(\Sigma y|M)}$$

$$= \frac{1}{1 + \exp R_m} \quad \ldots \ldots \quad (10)$$

where

$$R_m = \frac{1}{N_0}\int_0^T\left[-\sum_{u=1}^{p}(y_{mu})^2 - \sum_{v=1}^{q-p}(y_{sv} - s_v)^2 + \sum_{u=1}^{p}(y_{mu} - m_u)^2 + \sum_{v=1}^{q-p}(y_{sv})^2\right]dt$$

$$= \frac{1}{N_0}\int_0^T\left[\sum_{u=1}^{p}(m_u^2 - 2y_{mu}m_u) + \sum_{v=1}^{q-p}(2y_{sv}s_v - s_v^2)\right]dt \quad . \quad (11)$$

If mark is actually sent

$$y_{mu} = m_u + n_u$$

$$y_{sv} = n_v$$

where n_u and n_v represent noise. So that R_{mm}, the value of R_m when mark is actually sent, is given by

$$R_{mm} = \frac{1}{N_0}\int_0^T\left[\sum_{u=1}^{p}(-m_u^2 - 2m_u n_u) + \sum_{v=1}^{q}(2s_v n_v - s_v^2)\right]dt$$

$$= -\frac{\sum_{u=1}^{p} w_{mu} + \sum_{v=1}^{q-p} w_{sv}}{N_0} - \frac{2}{N_0}\int_0^T\left[\sum_{u=1}^{p}(m_u n_u) - \sum_{v=1}^{q-p}(s_v n_v)\right]dt$$

where w_{mu} and w_{sv} are respectively the energies of the signals in the uth mark and vth space branches given by

$$\left.\begin{aligned} w_{mu} &= \int_0^T m_u^2 dt \\ w_{sv} &= \int_0^T s_v^2 dt \end{aligned}\right\} \quad \ldots \ldots \quad (12)$$

The expression $\frac{2}{N_0}\int_0 m_u n_u dt$ has a Gaussian distribution of mean value zero and mean-square value $2w_{mu}/N_0$ by eqn. (5), and similarly for $\frac{2}{N_0}\int_0^T s_v n_v dt$. Since they arise from separate diversity branches the q distributions will be assumed to be independent, so that their sum is also a Gaussian distribution with mean value zero and mean-square value

$$\frac{2}{N_0}\left(\sum_{u=1}^{p} w_{mu} + \sum_{v=1}^{q-p} w_{sv}\right) = 4w_e/N_0 \quad . \quad . \quad . \quad (13)$$

In a practical diversity system the mark and space branches would be in pairs, with a single aerial feeding each pair. In these circumstances it is not strictly valid to assume that all the q distributions are independent; the effect of interaction between the mark and space signals is discussed in Section 3.5. The quantity w_e defined in eqn. (13) may be regarded as the effective signal energy per signal element. In the particular case in question, that of equal *a priori* probabilities of mark and space, the average total received energy per element is equal to w_e.

Let us identify as mark any signal for which

$$p(\varepsilon_m|y) > \tfrac{1}{2}$$

i.e. for which $\quad R_m < 0$

The probability of this inequality not being satisfied when mark is sent, i.e. the probability of error, is

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left[\frac{2w_e}{N_0}\left(\frac{N_0}{4w_e}\right)^{1/2}\right]$$

i.e.

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\,(w_e/N_0)^{1/2} \quad . \quad . \quad . \quad . \quad . \quad (14)$$

Since mark and space enter the work symmetrically, eqn. (14) also gives the probability of error when space is sent and hence gives the total error probability. The function is plotted in Fig. 1.

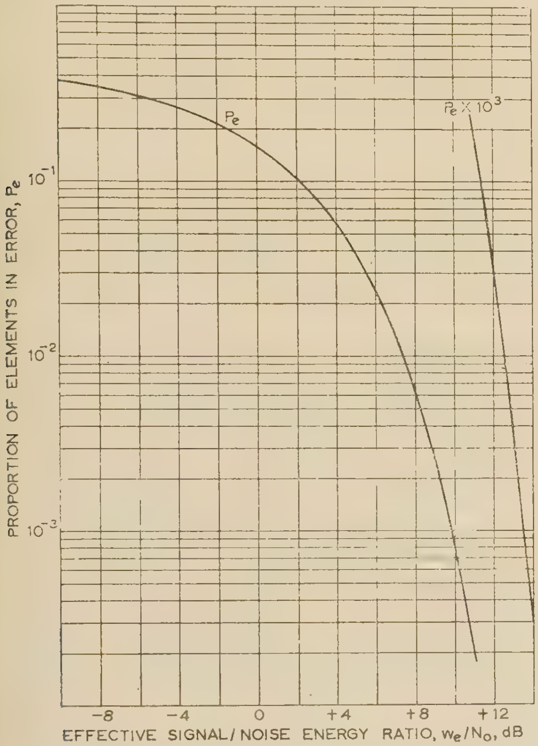


Fig. 1.—Error probability with steady signals.

A review of the results shows that the ideal diversity receiver combines the outputs of the various diversity branches. Practical diversity receivers usually select the largest output and ignore the rest; the loss entailed in this process is referred to again in Section 3.3 and discussed in more detail in a companion paper.[3] Eqn. (13) shows, in agreement with Kahn,[4] that the contributions of the various branches should be weighted in accordance with the signal energies in them. A mathematical basis for receiver design is contained in eqn. (11), which shows that in the ideal arrangement separate correlation processes are involved in the different diversity branches.

(3.2) Effective Signal-Energy Level Distribution

In a long-distance high-frequency system the signals in the individual diversity branches will vary in level, normally according to the Rayleigh distribution.[5] Thus the effective signal energy w_e, defined in Section 3.1, will also vary, in a manner dependent on the number of diversity branches, on the average energies in them and on the degree of correlation between their fading patterns. As the effective energy varies, so will the error liability; in order to determine the overall long-term error liability it is necessary to find the relative probabilities of various effective energy levels, i.e. the effective signal-energy density distribution, to multiply the figures so obtained by the error liabilities corresponding [eqn. (14)], and then to integrate over the full range of signal level. This Section is concerned with the probability density distribution of the effective signal energy.

The following assumptions are made:

(*a*) The fading is so slow relative to the speed of signalling that the signal power may be regarded as constant during any one signal element.

(*b*) The probability $p(w)$ that the signal-on energy in a branch is less than w is given by the Rayleigh formula

$$p(w) = 1 - \exp(-w/W) \quad . \quad . \quad . \quad . \quad (15)$$

where W is the mean energy in the branch in the signal-on condition.

(*c*) The mean signal-on energies of all branches are equal.

(*d*) The fadings in the different branches are uncorrelated.

Considering first the single-branch case, the distribution is given by the Rayleigh formula, eqn. (15), and the probability density is obtained by differentiation:

$$dp(w) = \frac{1}{W}\exp(-w/W)dw \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

$$= 0{\cdot}23\frac{w}{W}\exp(-w/W)d(10\log_{10} w)$$

so that the probability density per decibel is given by

$$\frac{dp(w)}{d(10\log_{10} w)} = 0{\cdot}23\frac{w}{W}\exp(-w/W) \quad . \quad . \quad (17)$$

It is convenient to express the density on a per-decibel basis for the purpose of computation.

In the case of dual diversity the total signal energy w is the sum of the signal energies, w_a and w_b, in the two branches:

$$w = w_a + w_b$$

By eqn. (16) the probability density for w_b is given by

$$dp(w_b) = \frac{1}{W}\exp(-w_b/W)dw_b$$

Writing $w - w_a$ for w_b in this expression, we have for the probability density of w as a function of w_a

$$dp(w)]_{w_a} = \frac{1}{W}\exp[-(w - w_a)/W]dw$$

For a total signal energy w the component w_a can have any value in the range $0 < w_a < w$, the probability density of w_a being given by

$$dp(w_a) = \frac{1}{W}\exp(-w_a/W)dw_a$$

so that the overall probability density for w is given by

$$dp(w) = dw\int_0^w \frac{1}{W}\exp[-(w - w_a)/W]\frac{1}{W}\exp(-w_a/W)dw_a$$

$$= \frac{w}{W^2}\exp(-w/W)dw \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

and the probability density per decibel for the case of dual diversity is

$$dp(w)/d(10\log_{10} w) = 0{\cdot}23(w/W)^2\exp(-w/W) \quad . \quad (19)$$

The triple-diversity case has been calculated similarly on the

basis of eqns. (16) and (18), giving the probability density per decibel

$$dp(w)/d(10 \log_{10} w) = \frac{0 \cdot 23}{2}(w/W)^3 \exp(-w/W) \quad . \quad (20)$$

and for q diversity branches

$$dp(w)/d(10 \log_{10} w) = \frac{0 \cdot 23}{(q-1)!}(w/W)^q \exp(-w/W) \quad (21)$$

Eqn. (21) gives the probability density per decibel of the sum of the signal-on energies in the q diversity branches, i.e. the sum of the energies in the mark branches when mark is sent plus the energies in the space branches when space is sent. The error probability for the diversity case, eqn. (14), has been worked out in terms of w_e, the effective value per signal element of the signal energy received, taking all the branches together.

This effective signal energy per element is, by definition [eqn. (13)], equal to half the sum of the signal-on energies, so that the probability density for the effective signal energy is identical with the probability density for the sum of the signal-on energies at the value $2w_e$. Substituting $w = 2w_e$ in the right-hand side of eqn. (21) we have

$$dp(w_e)/d(10 \log_{10} w_e) = \frac{0 \cdot 23}{(q-1)!}(2w_e/W)^q \exp(2w_e/W)$$
$$= \frac{0 \cdot 23}{(q-1)!}(w_e/W_0)^q \exp(w_e/W_0) \quad (22)$$

where W_0 is half the mean signal-on energy for one branch, i.e. the mean effective signal energy for one branch. The probability

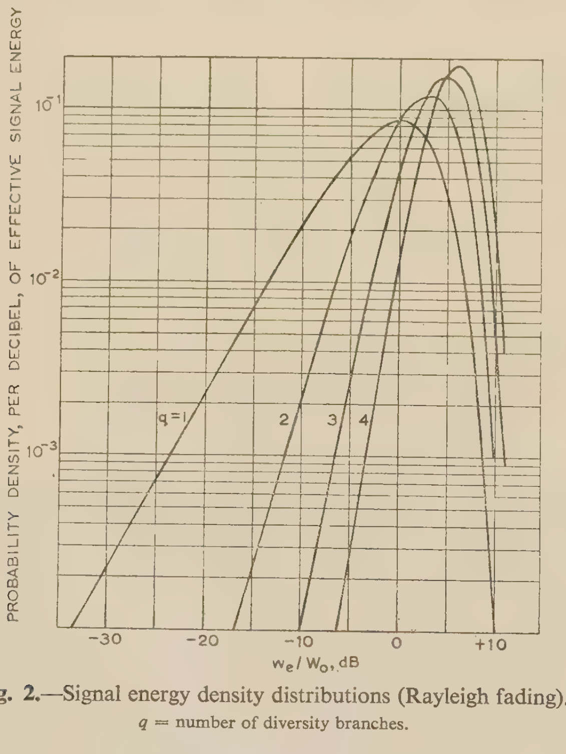


Fig. 2.—Signal energy density distributions (Rayleigh fading).
q = number of diversity branches.

densities per decibel are plotted in Fig. 2 for values of q of 1, 2, 3 and 4.

The assumption has been made above that the signal power may be regarded as constant during any one signal element. This assumption may not go far enough in practice, for the ideal performance is dependent on the prior knowledge of the amplitude of each signal element for both of the two possible signal conditions; this can only be derived in practice by measuring the amplitude of other elements occurring before or, with greater practical difficulty, after the element whose condition is to be determined. Thus, for maximum simplicity in amplitude forecasting the requirement is that the amplitude should not change significantly from one element to the next of the same type. Telegraph codes that permit a given condition to be maintained for an appreciable number of elements are at a disadvantage in this respect.

(3.3) Probability of Error with Fading Signals

The expressions for the probability of error when the total effective signal energy is w_e, eqn. (14), and for the probability density of w_e with q-fold diversity, eqn. (22), may be multiplied together to give the probability density of error, dP_{eq}, as a function of w_e:

$$dP_{eq}/d(10 \log_{10} w_e) = P_e dp(w_e)/d(10 \log_{10} w_e)$$
$$= [\tfrac{1}{2} - \tfrac{1}{2} \operatorname{erf}(w_e/N_0)^{1/2}]\frac{0 \cdot 23}{(q-1)!}(w_e/W_0)^q \exp(-w_e/W_0) \quad . \quad . \quad . \quad . \quad (23)$$

The total probability of error, in terms of the mean effective signal/noise energy ratio per branch, W_0/N_0, is obtained by the integration of eqn. (23); analytic methods appear to break down at this point and recourse has been had to numerical methods of

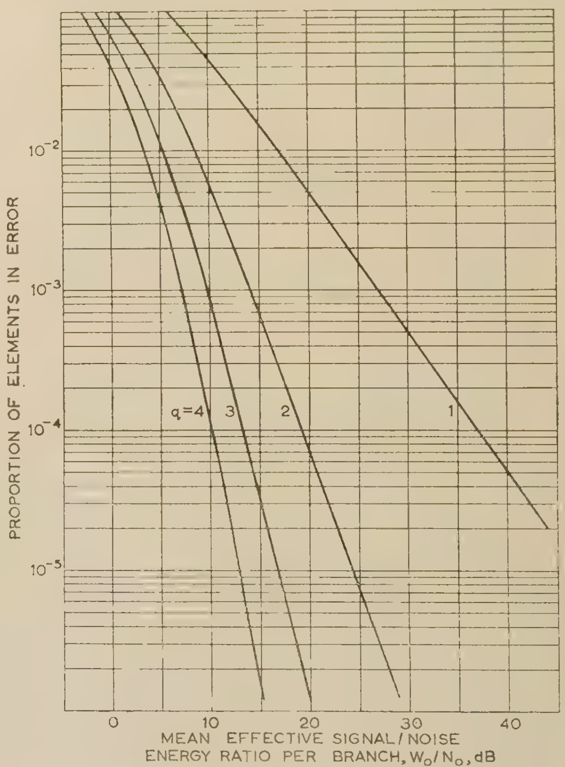


Fig. 3.—Error probability with Rayleigh-fading signals.
q = number of diversity branches.
(Mark and space equally likely to be sent.)

integration for the cases in which $q = 1$, 2, 3 and 4. The results are plotted in Fig. 3.

For some purposes it is more convenient to have relationships expressed in formulae rather than in graphs, and it is perhaps useful to record that the curves of Fig. 3 fit, within a fraction of a decibel, the formula[3]

$$P_{eq} = \frac{1}{2} \frac{q!}{(1 + dW_0/N_0)(2 + dW_0/N_0) \ldots (q + dW_0/N_0)} \quad (24)$$

where d is a factor dependent on q, and related to it thus:

For $q =$	1	2	3	4
$10 \log_{10} d =$	0	0·7	1·3	1·9 dB

The factor d is of some significance, for it represents the loss involved in diversity by selection with the limiter-discriminator type of receiver[3] (see Appendix 8, page 140).

Some typical plots of error density $dP_{eq}/d(10 \log_{10} w_e)$ as a function of w_e/W_0 with W_0/N_0 as a parameter are given in Figs. 4,

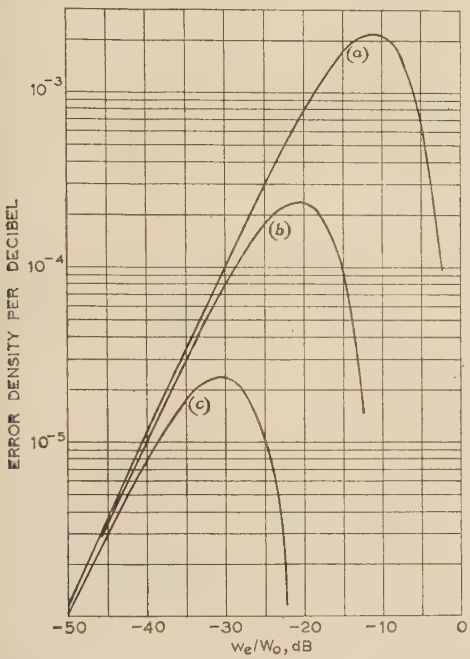


Fig. 4.—Error density distribution (Rayleigh fading).
No diversity ($q = 1$).
(a) $W_0/N_0 = +12$ dB; $P_{eq} = 2{\cdot}9 \times 10^{-2}$.
(b) $W_0/N_0 = +22$ dB; $P_{eq} = 3{\cdot}1 \times 10^{-3}$.
(c) $W_0/N_0 = +32$ dB; $P_{eq} = 3{\cdot}1 \times 10^{-4}$.

5 and 6 for $q = 1$, 2 and 4 respectively. These graphs demonstrate how the fading range that is of importance in the incidence of errors diminishes rapidly as the number of diversity branches increases. Thus, taking total error rates of the order of 1 in 10 000 elements, fading worse than -10 dB relative to the mean is so rare as to make a negligible contribution to the total errors with quadruple diversity, whereas fading of -40 dB matters in the no-diversity case. The corresponding figure for dual diversity is about -20 dB.

(3.4) Unequal *a priori* Probabilities of Mark and Space

The analysis so far has, for the sake of simplicity, been restricted to the case of equal *a priori* probabilities of mark and space. Telegraph codes in general use approximate to this condition, but it is desirable to know whether the differences that do exist significantly affect the error liability.

In the general case λ and μ in eqn. (1) are unequal, and in place of eqn. (10) the existence probability of mark is given by

$$p(\varepsilon_m|y) = \frac{1}{1 + \mu p(\Sigma y|S)/\lambda p(\Sigma y|M)}$$

$$= \frac{1}{1 + \exp(R_m + a)} \quad . \quad . \quad . \quad . \quad (25)$$

where R_m is defined by eqn. (11) as before and a is given by

$$a = \log_\varepsilon (\mu/\lambda) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (26)$$

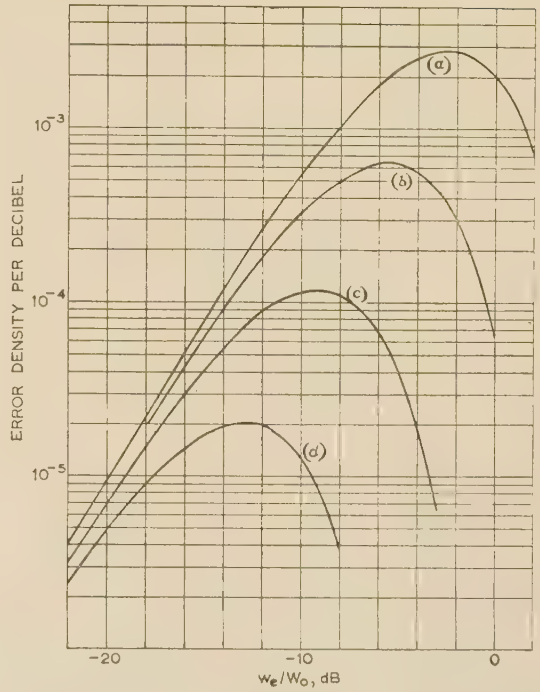


Fig. 5.—Error density distribution (Rayleigh fading).
Dual diversity ($q = 2$).
(a) $W_0/N_0 = +6$ dB; $P_{eq} = 2{\cdot}4 \times 10^{-2}$.
(b) $W_0/N_0 = +10$ dB; $P_{eq} = 5{\cdot}6 \times 10^{-3}$.
(c) $W_0/N_0 = +14$ dB; $P_{eq} = 1{\cdot}05 \times 10^{-3}$.
(d) $W_0/N_0 = +18$ dB; $P_{eq} = 1{\cdot}8 \times 10^{-4}$.

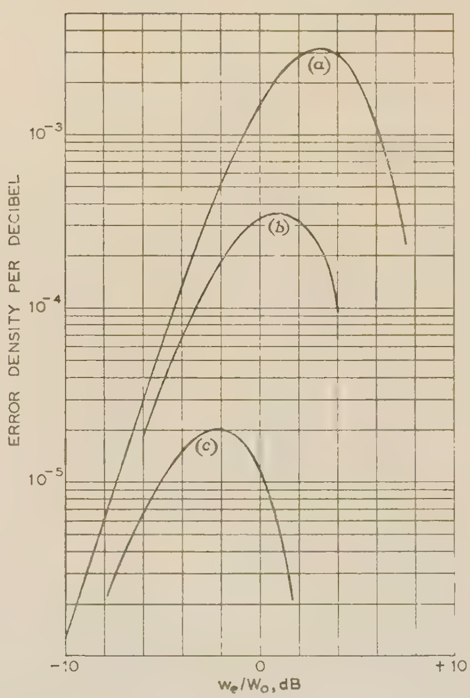


Fig. 6.—Error density distribution (Rayleigh fading).
Quadruple diversity ($q = 4$).
(a) $W_0/N_0 = +2$ dB; $P_{eq} = 1{\cdot}8 \times 10^{-2}$.
(b) $W_0/N_0 = +6$ dB; $P_{eq} = 2{\cdot}1 \times 10^{-3}$.
(c) $W_0/N_0 = +10$ dB; $P_{eq} = 1{\cdot}1 \times 10^{-4}$.

Let us identify as mark any signal for which

$$p(\varepsilon_m|y) > \tfrac{1}{2}$$

i.e. $$R_m < -a$$

Thus, in place of eqn. (14) the probability of error when mark is sent, P_{em}, is given by

$$P_{em} = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left[\left(\frac{2w_e}{N_0} - a\right)(N_0/4w_e)^{1/2}\right]$$

$$= \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\,[(w_e/N_0)^{1/2} - \tfrac{1}{2}a(N_0/w_e)^{1/2}]$$

Similarly, when space is sent

$$P_{es} = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\,[(w_e/N_0)^{1/2} + \tfrac{1}{2}a(N_0/w_e)^{1/2}]$$

Hence for the overall probability of error, P_e, we have

$$\begin{aligned} P_e &= \lambda P_{em} + \mu P_{es} \\ &= \tfrac{1}{2}(\lambda + \mu) - \tfrac{1}{2}\lambda\,\mathrm{erf}\,[(w_e/N_0)^{1/2} - \tfrac{1}{2}a(N_0/w_e)^{1/2}] \\ &\quad - \tfrac{1}{2}\mu\,\mathrm{erf}\,[(w_e/N_0)^{1/2} + \tfrac{1}{2}a(N_0/w_e)^{1/2}] \quad (27) \end{aligned}$$

The probability of error is still controlled by the effective signal energy per element, w_e, although this is no longer the same as the average signal energy per element, since we are dealing with the case of unequal *a priori* probabilities of mark and space. For the average signal energy per element, w_{av}, we have

$$\frac{w_e}{w_{av}} = \frac{\tfrac{1}{2}\Sigma w_{mu} + \tfrac{1}{2}\Sigma w_{sv}}{\lambda\Sigma w_{mu} + \mu\Sigma w_{sv}}$$

Two cases are of special interest. First, with two-tone and similar signals, in the absence of selective fading, it can be assumed that

$$\Sigma w_{mu} = \Sigma w_{sv}$$

so that $$w_e/w_{av} = (\tfrac{1}{2} + \tfrac{1}{2})/(\lambda + \mu) = 1$$

In the case of on-off signalling, $w_{sv} = 0$, so that

$$w_e/w_{av} = 1/2\lambda$$

If λ is small compared with unity the effective signal energy per element is much larger than the average signal energy per element.

Curves showing the error liabilities as functions of the *average* signal/noise energy ratio for $\lambda = 0{\cdot}25$ and $\lambda = 0{\cdot}1$, and for $\Sigma w_{sv} = \Sigma w_{mu}$ and $\Sigma w_{sv} = 0$, are plotted in Fig. 7 for comparison with the case already obtained in which $\lambda = 0{\cdot}5$. The curves for the case of equal mark and space energies lie close to the $\lambda = 0{\cdot}5$ curve, particularly at reasonably low error rates. Figures which have been calculated for the important 4-mark : 3-space error-detecting codes lies so close to the 1 : 1 case ($\lambda = 0{\cdot}5$) that they have not been separately plotted; evidently all the fading and diversity results obtained for the 1 : 1 case can be applied to the 4 : 3 case without significant error.

The curves in Fig. 7 indicate that, considered on an energy basis, useful gains can be obtained by the use of highly unsymmetrical codes with on-off signalling. It has to be remembered, however, that such codes are comparatively inefficient, so that the reduced element-error liability is offset to some extent by the increased number of elements required per character. Also, with on-off keying a frequency-diversity advantage can be obtained only by the use of carrier waves modulated at audio frequency (A2 emissions), a practice that is usually deprecated because of the resulting bandwidth occupancy, though this is not necessarily any worse than that of a wide-shift two-tone or frequency-shift signal.

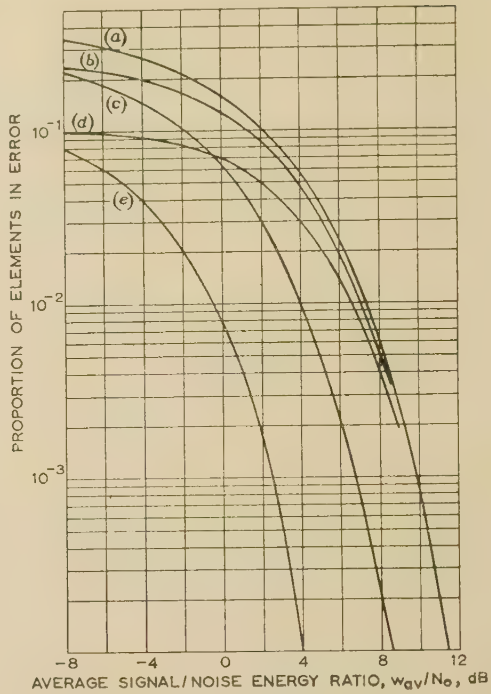


Fig. 7.—Error probability with steady signals of various *a priori* probabilities of mark.

Curve	*A priori* probability of mark	Energies per element relative to average energy per element		
		Mark	Space	Effective
(*a*)	0·5	*g*	2 − *g*	1
(*b*)	0·25	1	1	1
(*c*)	0·25	4	0	2
(*d*)	0·1	1	1	1
(*e*)	0·1	10	0	5

(Note: *g* can have any value between 0 and 2.)

(3.5) Effect of Correlation between Mark and Space Signals

It was assumed in Section 3.1 that the mark and space detectors of a two-tone or frequency-shift receiver were fed with entirely independent signals and noise. In a practical radio or line system a single aerial or line would handle both mark and space signals and the noise would, of course, be common. The action of the ideal receiver under these conditions will now be examined.

If mark and space detectors are fed in pairs from common aerials the number of space branches will be the same as the number of mark branches and the two members of a pair will be fed with identical waveforms. Thus, in terms of the symbols of Section 3.1, we have $q = 2p$ and $y_{mu} = y_{sv} = y_u$, say. The *a priori* probabilities of mark and space are again assumed to be equal and in place of eqn. (11) we have

$$R_m = \frac{1}{N_0}\int_0^T \sum_{u=1}^{p}(m_u^2 - 2y_u m_u + 2y_u s_u - s_u^2)dt \quad . \quad (28)$$

If mark is actually sent

$$y_u = m_u + n_u$$

and the value of R_m when mark is sent is

$$R_{mm} = \frac{1}{N_0}\int_0^T \sum_{u=1}^{p}(-m_u^2 - 2m_un_u + 2m_us_u + 2s_un_u - s_u^2)dt$$

$$= -\frac{1}{N_0}\sum_{u=1}^{p}\left(w_{mu} + w_{su} - 2\int_0^T m_us_udt\right)$$

$$-\frac{2}{N_0}\int_0^T \sum_{u=1}^{p}(m_u - s_u)n_udt$$

$$= -\frac{1}{N_0}\sum_{u=1}^{p}\left(w_{mu} + w_{su} - 2\int_0^T m_us_udt\right) - \sum_{u=1}^{p} z_u \quad (29)$$

where

$$z_u = \frac{2}{N_0}\int_0^T (m_u - s_u)n_udt \quad . \quad . \quad . \quad . \quad (30)$$

The function z_u has a Gaussian distribution of mean value zero and mean-square value given by

$$\overline{z_u^2} = \frac{2}{N_0}\int_0^T (m_u - s_u)^2dt$$

$$= \frac{2}{N_0}\left(w_{mu} + w_{su} - 2\int_0^T m_us_udt\right) \quad . \quad . \quad (31)$$

The p distributions may be assumed to be independent since they arise from different aerials, so that

$$\sum_{u=1}^{p}\overline{z_u^2} = \frac{2}{N_0}\sum_{u=1}^{p}(w_{mu} + w_{su} - 2\int_0^T m_us_udt) \quad . \quad (32)$$

An error will arise if $R_{mm} > 0$ and the probability of this occurring is given by

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left[\frac{1}{2N_0}\sum_{u=1}^{p}\left(w_{mu} + w_{su} - 2\int_0^T m_us_udt\right)\right]^{1/2}$$

$$= \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left[\frac{w_e}{N_0}\left(1 - \frac{1}{w_e}\int_0^T \sum_{u=1}^{p} m_us_udt\right)\right]^{1/2}$$

$$= \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left[\frac{w_e}{N_0}(1 - \rho)\right]^{1/2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (33)$$

where

$$\rho = \frac{1}{w_e}\int_0^T \sum_{u=1}^{p} m_us_udt \quad . \quad . \quad . \quad . \quad . \quad (34)$$

The signal/noise ratio is therefore modified by a term involving the cross-correlation of the mark and space signals. If the correlation is negative the performance of the system is improved and the largest possible improvement, namely 3 dB, is obtained if $s_u = -m_u$, when we have

$$\rho = \frac{1}{w_e}\int_0^T \sum_{u=1}^{p}(-m_u^2)dt = -1$$

so that

$$P_e = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{erf}\left(\frac{2w_e}{N_0}\right)$$

The ideal modulation waveform for telegraphy is square. In practical systems some slowing down of transitions and rounding of corners result from limitations of transmitted bandwidth, but these effects are not severe in high-frequency radiotelegraph practice. It is therefore reasonable to continue the analysis on the basis of square-wave modulation, and the undistorted mark and space signals corresponding to an element starting at time $t = 0$ in a two-tone system may be written thus:

$$m_u = v_{mu}\cos\left[2\pi(f_c + \tfrac{1}{2}f_d)t + \phi_{mu}\right] \qquad 0 < t < T$$
$$s_u = v_{su}\cos\left[2\pi(f_c - \tfrac{1}{2}f_d)t + \phi_{su}\right] \qquad 0 < t < T$$

where v_{mu} and v_{sv} = Peak voltages.

f_d = Frequency difference between mark and space.

f_c = Mean or centre frequency.

ϕ_{mu} and ϕ_{su} = Phase angles.

Thus, substituting in eqn. (34), we have

$$\rho = \frac{1}{2w_e}\sum_{u=1}^{p} v_{mu}v_{su}\left[\frac{1}{4\pi f_c}\sin(4\pi f_ct + \phi_{mu} + \phi_{su}) + \frac{1}{2\pi f_d}\sin(2\pi f_dt + \phi_{mu} - \phi_{su})\right]_0^T$$

Provided that at the detector the carrier frequency is very large compared with the frequency difference, the carrier frequency term in the above expression may be neglected and we have

$$\rho = \frac{1}{2w_e}\sum_{u=1}^{p}\frac{v_{mu}v_{su}}{2\pi f_d}\left[\sin(2\pi f_dT + \phi_{mu} - \phi_{su}) - \sin(\phi_{mu} - \phi_{su})\right] \quad . \quad . \quad . \quad . \quad (35)$$

Eqn. (35) shows that the correlation depends on the relative magnitudes, phases and frequency differences of the undistorted mark and space signals in each aerial. The correlation can only be significant if the undistorted mark and space signals have similar amplitudes in each aerial that is contributing significantly to the total signal energy, and to take the extreme case we may put $v_{mu} = v_{su}$ so that

$$\sum_{u=1}^{p} v_{mu}v_{su}T = 2w_e$$

Then, assuming that the phase difference $\phi_{mu} - \phi_{su}$ is the same for each aerial, we have

$$\rho = \frac{1}{2\pi f_dT}\left[\sin(2\pi f_dT + \phi_m - \phi_s) - \sin(\phi_m - \phi_s)\right]$$

Limiting values are given by putting

$$\phi_m - \phi_s = -\left(\pi f_dT - \frac{\pi}{2} \pm \frac{\pi}{2}\right)$$

so that

$$\rho = \pm\frac{\sin(\pi f_dT)}{\pi f_dT} \quad . \quad . \quad . \quad . \quad . \quad (36)$$

If the signal arises from the frequency modulation of an oscillator, phase-continuity considerations require that $\phi_m - \phi_s = 0$, and the correlation in this case is given by

$$\rho = \frac{\sin(2\pi f_dT)}{2\pi f_dT} \quad . \quad . \quad . \quad . \quad . \quad (37)$$

The effect of cross-correlation between mark and space signals is that the effective signal energy is multiplied by a factor $(1 - \rho)$, as shown in eqn. (33). The values of this factor for the extreme case given by eqn. (36) and for frequency modulation given by eqn. (37) are shown in Fig. 8. Even in the extreme case substantial effects due to cross-correlation appear only with values of f_dT appreciably less than unity. The maximum possible effect is less than 0·6 dB for f_dT values greater than 1·8. It will be

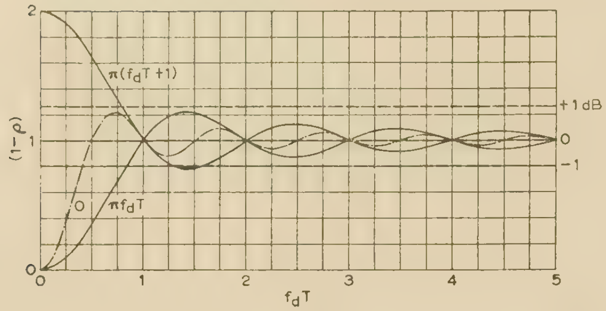


Fig. 8.—Correlation effects between mark and space signals.
Values of ϕ_s—ϕ_m shown against curves.

noticed that there is no cross-correlation at integral values of f_dT, and at half-integral values also with frequency modulation. On practical high-frequency radiotelegraph channels using frequency modulation or two-tone signalling, values of f_dT commonly lie in the range 2·4 to 10. Evidently the extreme errors involved in ignoring cross-correlation in such systems are small. Furthermore, the extreme conditions are unlikely to be important in high-frequency practice; such effects as selective fading and, in the case of two-tone signals, uncontrolled relative phasing of the mark and space signals, will reduce the strength of the correlation. There is therefore good justification for ignoring the correlation in these cases.

The condition in which $f_dT = 0$, $\phi_s - \phi_m = \pi$, which gives an improvement of 3 dB, is of course the case of phase modulation with deviation of $\pm\frac{1}{2}\pi$. It is attractive for line systems or for radio systems that are not given to selective fading. If suitably selective fading prevails or can be arranged to occur, frequency-modulation or two-tone signalling is to be preferred for the sake of the frequency diversity advantage that it can give. This subject is discussed in another paper.[6]

(4) CHARACTER-ERROR LIABILITY

The discussion so far has been in terms of the probability of wrong elements. In practical telegraph codes, elements are treated in groups, each representing a character, i.e. a letter, figure, space, punctuation mark, or other printer function, so that errors in reception manifest themselves as wrong characters in the printed message; operational assessments of the performance of telegraph systems are therefore made in terms of the proportion of wrong characters. Many different telegraph codes are in use, and it would be inappropriate to consider them all here, but they can be broadly divided into two classes, (*a*) codes that are substantially free from redundancy, and (*b*) 'protected' codes containing a substantial measure of redundancy which is used to detect, or correct, errors. In this Section two codes, typical of each type, will be treated, namely the 5-unit code, in which each character comprises five telegraph elements giving 32 combinations, and the 4-mark/3-space (7-unit) error-detecting code, in which each character comprises seven telegraph elements (four marks and three spaces), giving 35 combinations. The assumption is made that the element errors are randomly incident.

If P_e is the probability of element error and r is the number of elements forming a character, the probability of receiving a character correctly, P_{c0r}, is given by

$$P_{c0r} = (1 - P_e)^r$$

The probability of any one element in a character being wrong and the rest being right, P_{c1r}, is given by

$$P_{c1r} = rP_e(1 - P_e)^{r-1}$$

and for the probability of any number, b, of elements being wrong we have

$$P_{cbr} = \frac{r!}{b!(r-b)!}P_e^b(1 - P_e)^{r-b} \quad . \quad . \quad . \quad (38)$$

With the unprotected 5-unit code any element error must cause a character error, which will be undetectable except in relation to its context. Thus the character undetected-error rate, P_{cu5}, is related to the element-error probability by the equation

$$P_{cu5} = 1 - (1 - P_e)^5 \quad . \quad . \quad . \quad . \quad (39)$$

By applying this equation to the element-error probabilities given in Fig. 3 the character error rate for any value of W_0/N_0 can be determined. The signal/noise energy ratio W_0/N_0 is, of course, related to the energy in a signal element, but since we are now dealing with characters it is more logical to work in terms of the corresponding mean effective energy in a character, W_{co}, given by

$$W_{co} = rW_0$$

In the present case $r = 5$ and an adjustment of 7 dB is necessary in the signal/noise energy-ratio scale. The results of applying these processes to the cases of 1, 2 and 4 diversity branches are shown in the solid curves in Fig. 9.

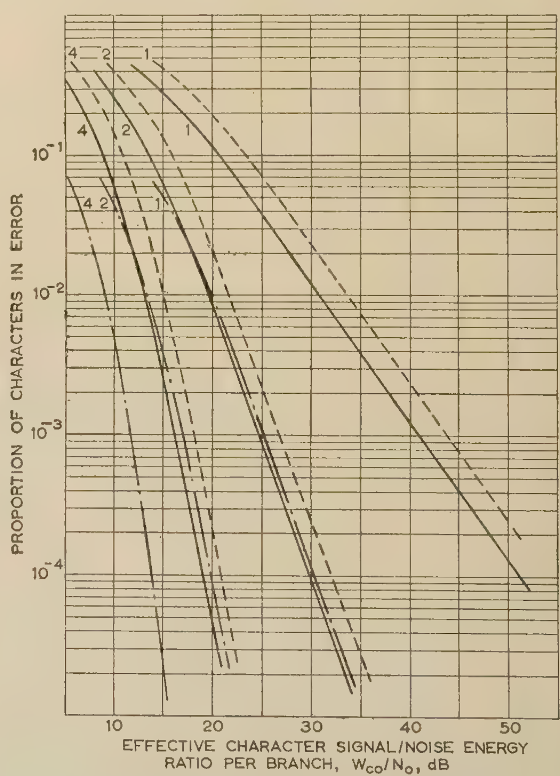


Fig. 9.—Character-error probabilities of 5-unit and 7-unit error-detecting codes with Rayleigh-fading signals.

Number of diversity branches shown against curves.
——— 5-unit code.
– – – 7-unit code detected errors.
—·— 7-unit code undetected errors.

With the 7-unit code it is necessary to distinguish between detectable and undetectable errors. Thus, any single element error must disturb the 4/3 relationship and is therefore detectable. Similarly, half the possible double errors are detectable, but the others are equivalent to transpositions and are undetectable. All triple errors are detectable. This is as far as it is necessary to go, since, for reasonable undetected-error rates, higher numbers

of element errors in a character contribute negligibly to the total. Thus, for the undetected and detected character error rates, P_{cu7} and P_{cd7}, we have

$$P_{cu7} \simeq \tfrac{1}{2}P_{c27}$$

$$\simeq \frac{1}{2}\frac{7.6}{1.2}P_e^2(1 - P_e)^5 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (40)$$

$$P_{cd7} \simeq P_{c17} + \tfrac{1}{2}P_{c27} + P_{c37}$$

$$\simeq 7P_e(1 - P_e)^6\left[1 + \frac{3}{2}\left(\frac{P_e}{1 - P_e}\right) + 5\left(\frac{P_e}{1 - P_e}\right)^2\right] \quad (41)$$

The application of these equations to the element-error probabilities of Fig. 3, for the cases of 1, 2 and 4 diversity branches, and an adjustment of 8·45 dB in the signal/noise energy scale, corresponding to $r = 7$, give the detected and undected character-error probabilities shown respectively in the dotted and chain-dotted curves of Fig. 9.

From a comparison of the various curves of Fig. 9 the interesting point emerges that, so far as the undetected error rate is concerned, changing from the 5-unit to the 7-unit code is roughly equivalent to doubling the number of diversity branches; inspection of eqns. (24), (39) and (40) reveals the same point. This is, of course, not the whole story, for the detected errors in the 7-unit case have to be cleared by reference back to the sending end, and this involves some loss of capacity on both 'go' and 'return' channels. When automatic reference-back equipment is used (auto-RQ working)[7] the loss becomes appreciable only when the error rate becomes very large; conditions that with the unprotected code would give a character error rate of 1 in 1 000 and that would therefore, with the operational methods commonly employed at present, require every message to be sent twice (slips-twice working—ZST) so halving the channel capacity, would reduce the capacity by only a few per cent in auto-RQ working using the 7-unit code.

(5) DISCUSSION AND CONCLUSIONS

The noisy-signal performance of an ideal receiver provides an absolute basis for assessing the performance of practical equipment, which can be described in terms of the amount by which it falls short of the ideal. It can conveniently be expressed as the ratio of the energy required to produce a given error rate in the practical equipment to that required for the same error rate in the ideal receiver, the noise being the same in both cases. This measure of the imperfection of equipment seems likely to be of considerable practical importance and a name for it is needed. Since it is largely a function of the method of demodulation, the term 'demodulation factor' is tentatively suggested. Alternatively, to indicate its connection with noise, 'demodulation noise factor' might be used.

The demodulation factor of an equipment will, in general, vary with the speed of signalling and the error rate. Sometimes, however, the variation with error rate may be so small that a single value applies over a wide range. The interpretation of results is then much simplified. From this point of view fading has a useful effect in reducing the degree in which demodulation factor depends upon error rate. Thus the exponential relation between error rate and signal/noise ratio that characterizes[3] the performance of frequency-shift telegraph receivers of the conventional limiter/frequency-discriminator type differs appreciably from the error-function relation of the ideal case. On fading signals, however, the demodulation factor varies by only a fraction of a decibel over the range of error rate that is of practical interest, and for most purposes the variation can be ignored.

A knowledge of the demodulation factor is useful in development work, for it indicates the amount of improvement theoretically possible. Another, and more direct, guide to development is provided in eqn. (11), which may be regarded as a mathematical specification for the ideal receiver. In effect, the received signals in each diversity branch are correlated with perfect signals produced locally. It will be noticed that absence of signal in a branch is as significant as presence, the branch outputs being, of course, of opposite polarity in the two cases. The outputs of the diversity branches are weighted according to the signal energies they transmit and are then combined. The extent to which the processes of eqn. (11) should be reproduced in practical equipment will depend on the degree of difficulty involved and on the importance of having a good demodulation factor in any particular case. As an example of practical difficulty the question of carrier phase may be cited. Ideal demodulation involves making use of knowledge of carrier phase, and this would involve some complication even in the simple case of line telegraphy; the complication would be severe in the h.f. radio case, where random variation of the carrier phase always accompanies the fading. An example of the use of eqn. (11) as a guide in practical design is contained in a companion paper.[6]

Two further points concerning diversity working arise from the analysis. First, the error density distributions over the fading range, Figs. 5 and 6, show that deep fading contributes insignificantly to the total errors when diversity reception is employed. Thus it is not necessary, when specifying diversity receiving equipment, to call for good performance under very deep fading; it may be undesirable to do so, for an unnecessary constraint on the design in this respect may be prejudicial to the cost or to the performance in other respects. Secondly, the analysis shows that, so far as the incidence of undetected character errors is concerned, changing from an unprotected 5-unit code to a 4/3 error-detecting code is roughly equivalent to doubling the number of diversity branches. Knowledge of this equivalence should be useful when considering alternative ways of improving a system in which noise is a limiting factor.

Cross-correlation effects between the mark and space signals of two-tone or frequency-modulation systems are not significant under normal conditions of long-distance high-frequency radiotelegraphy. They have been ignored in the main analysis. In the absence of selective fading, however, some improvement in performance, due to cross-correlation, results from the use of frequency differences between mark and space that are small relative to the speed of signalling, provided that the signals are appropriately phased. The maximum improvement, of 3 dB, is obtained with the well-known $\pm\pi/2$ phase modulation; this case is potentially of interest in line telegraphy or low-frequency radiotelegraphy. If selective fading prevails, larger improvements can be obtained[6] by using a suitably large mark-space frequency difference and taking advantage of frequency diversity.

The study here outlined was undertaken primarily to gain a better understanding of some of the problems of long-distance synchronous radiotelegraphy in the 4–30 Mc/s band, and the assumptions that have been made about the characteristics of fading are appropriate to this field. It is possible that they may be valid in other applications. In any event the method of attack should be generally applicable, though fresh assumptions about fading will lead to new density distributions for the effective signal energy, necessitating new integrations of eqn. (23). The mathematical specification for the ideal noisy-signal diversity receiver, eqn. (11), applies irrespective of the amplitude distribution of the fading signal, provided only that the fading within an element can be neglected. It is therefore hoped that the study will prove useful in various fields, including the important one of microwave pulse communication.

(6) ACKNOWLEDGMENTS

Acknowledgment is made to the Engineer-in-Chief of the Post Office and to the Controller of H.M. Stationery Office for permission to publish the paper.

(7) REFERENCES

(1) WOODWARD, P. M., and DAVIES, I. L.: 'Information Theory and Inverse Probability in Telecommunication', *Proceedings I.E.E.*, Paper No. 1239 R, March, 1952 (**99**, Part III, p. 37).

(2) DAVIES, I. L.: 'On Determining the Presence of Signals in Noise', *ibid.*, Paper No. 1256 R, March, 1952 (**99**, Part III, p. 45).

(3) LAW, H. B.: 'The Signal/Noise Performance Rating of Receivers for Long-Distance Radiotelegraph Systems using Frequency Modulation' (see page 124).

(4) KAHN, L. R.: 'Ratio Squarer', *Proceedings of the Institute of Radio Engineers*, **42**, p. 1704 (November, 1954).

(5) RATCLIFFE, J. A.: 'Diffraction from the Ionosphere and the Fading of Radio Waves', *Nature*, **162**, 4105, p. 9 (3 July, 1948).

(6) ALLNATT, J. W., JONES, E. D. J., and LAW, H. B.: 'Frequency Diversity in the Reception of Selectively Fading Binary Frequency-Modulated Signals, with special reference to Long-Distance Radiotelegraphy' (see page 98).

(7) HAYTON, T., HUGHES, C. J., and SAUNDERS, R. L.: 'Telegraph Codes and Code Convertors', *Proceedings I.E.E.*, Paper No. 1585 R, November, 1953 (**101**, Part III, p. 137).

(8) APPENDIX*

Comparison of Diversity Methods

In Section 3.3 of the present paper and in Section 3.1 of Paper No. 2103 (page 124), comparisons are made between the diversity improvements given by the ideal receiver and by the conventional limiter-discriminator receiver. These receivers differ in two respects, for they use different methods of diversity, ideal combination and selection of the most active path, and they have different steady-signal characteristics, the error-function formula [eqn. (14) of this paper] and the exponential formula [eqn. (2) of Paper No. 2103]. This distinction was not drawn in the papers as originally printed.

It is of interest to determine the fading-signal performance of a receiver which uses selection diversity, but which is ideal in all other respects. The error liability of such a receiver may be calculated by replacing the signal-energy density component of eqn. (23) by the corresponding component for selection diversity, derived from Section 3.1 of Paper No. 2103. Thus for the probability density, per decibel, of effective signal energy, w_e, we have in place of eqn. (23)

$$dp(w_e)/d(10\log_{10} w_e) = 0\cdot 23q\frac{w_e}{W_0}\exp\left(-\frac{w_e}{W_0}\right)\left[1-\exp\left(-\frac{w_e}{W_0}\right)\right]^{q-1} \quad (42)$$

giving the probability density of error, with q diversity branches operating on a selection basis, as follows:

$$dP_{eq}/d(10\log_{10} w_e) = \left[\frac{1}{2}-\frac{1}{2}\operatorname{erf}\left(\frac{w_e}{N_0}\right)^{1/2}\right] \times 0\cdot 23q\frac{w_e}{W_0}\exp\left(-\frac{w_e}{W_0}\right)\left[1-\exp\left(-\frac{w_e}{W_0}\right)\right]^{q-1} \quad (43)$$

* The Appendix was received 24th January, 1957.

The total probability of error may be calculated as before by numerical integration. This has been done for element error rates in the 1 in 10^4 region, with the results shown in the first row of Table 1. The results are expressed as a diversity combination loss, being the amount in decibels by which the diversity

Table 1

DIVERSITY COMBINATION LOSS IN DECIBELS

Type of receiver	Number of diversity branches		
	2	3	4
Selection diversity, but otherwise ideal	1·5	2·5	3·3
Selection diversity plus exponential steady-signal characteristic	0·7	1·3	1·9
Voltage combination, but otherwise ideal	0·6	(0·9)	(1·0)

improvement falls short of the ideal. The signal energy density distributions are very similar in shape to the ideal ones, the main difference being a lateral displacement which varies by a fraction of a decibel over the range of w_e/W_0 that is of practical interest. This suggests that the values of diversity combination loss calculated for the error rate 1 in 10^4 should be reasonably accurate for other error rates up to, say, 1 in 10^2. When the mean signal/noise ratio is high, greatly simplified expressions may be used for the probability densities of signal energy in eqns. (23) and (43), and these lead to the following approximation for the diversity combination loss, L_s say, arising in selection diversity in q diversity branches:

$$L_s \simeq 10\log_{10}(q!)^{1/q} \text{ decibels} \quad . \quad . \quad . \quad . \quad (44)$$

The second line of Table 1 repeats for comparison purposes the diversity combination losses for the case, studied in Paper No. 2103, of a receiver with an exponential steady-signal characteristic and selection diversity. This applies accurately for error rates in the region of 1 in 10^4.

Though signals should ideally be combined on a power or energy basis, it is much simpler[6] to combine them on a voltage basis, so that the corresponding diversity combination loss is of some interest. For dual diversity the probability density of effective signal energy is given by the following formula, due allowance having been made for the fact that the noises in the two branches also add, this time on a power basis:

$$dp(w_e)/d(10\log_{10} w_e) = \exp\left(-\frac{w_e}{W_0}\right) \times\left[\left(\frac{2w_e}{W_0}\right)^{1/2}\exp\left(-\frac{w_e}{W_0}\right)+1\cdot 25\left(\frac{2w_e}{W_0}-1\right)\operatorname{erf}\left(\frac{2w_e}{W_0}\right)^{1/2}\right] \quad (45)$$

This distribution also is very similar in shape to the corresponding ideal one, with a lateral shift corresponding to a diversity combination loss of 0·6 dB; this has been entered in the third line of the Table. The full formulae for larger numbers of diversity branches have not been obtained, but here again simplified expressions may be used when the mean signal/noise ratio is high, and these lead to an approximation for the diversity combination loss, L_v, say, in the case of voltage combination of q branches:

$$L_v \simeq 10\log_{10}\left(2q\{(q-1)!/2[(2q-1)!]\}^{1/q}\right) \text{ decibels} \quad (46)$$

Figures calculated from this formula have been inserted, in brackets, in the Table.

[The discussion on the above paper will be found on page 147.]

621.396.324 : 621.317.34

Paper No. 2259 R
Nov. 1956

LABORATORY TEST EQUIPMENT FOR SYNCHRONOUS REGENERATIVE RADIOTELEGRAPH SYSTEMS

By C. G. HILTON, H. B. LAW, B.Sc.Tech., Associate Member, F. J. LEE, Associate Member, and F. A. W. LEVETT.

(*The paper was first received* 14*th August, and in revised form* 9*th October*, 1956. *It was published in November*, 1956, *and was read before the* RADIO AND TELECOMMUNICATION SECTION 14*th November*, 1956.)

SUMMARY

Traditional methods of testing radiotelegraph equipment, in terms of telegraph distortion, give results that are difficult to interpret operationally. In regenerative systems, however, the occurrence of errors in the printed copy is closely related to the element-error liability, and this fact has led to the development of error-counting test equipment in which signals passed through a system under test are compared with perfect signals direct from the signal source. The equipment covers a wide range of telegraph speeds, is easy to use and silent in operation, and it has proved to be a powerful tool in investigations of system behaviour and in development work. The suggestion is made that performance of regenerative systems ought in future to be specified in terms of element-error liability instead of telegraph distortion.

(1) INTRODUCTION

The development of communication systems and the specification of their component equipments depend greatly on effective methods of measuring performance. A good method will be easy to use, suitably accurate and reasonable in its requirements in test apparatus, and the effectiveness of the equipment in normal use will be simply related to the measured performance.

The performance of telegraph systems is usually described in terms of telegraph distortion, which is the ratio of the displacements of transitions from their correct instants in time to the duration of a telegraph element. Telegraph distortion is easily measured, and its use as a basis for specification facilitates the subdivision of tolerances between a number of links operating in tandem; this is very convenient in line-telegraphy practice. In long-distance radiotelegraphy, however, the question of subdividing tolerances hardly arises, for radio links are not normally operated in tandem without regeneration, and the distortion arising in any line tails associated with the radio link should be very small. Thus the advantage of easy subdivision is of little avail. Furthermore, radio noise, fading and interference are apt to cause additional transitions ('splits' and 'extras': see Fig. 1)—events that cannot be described in terms of telegraph distortion.

Most of the important point-to-point radiotelegraph services operated by the Post Office are of the direct-printing synchronous kind, the received telegraph signals being regenerated by sampling at the centre of each telegraph element. The operational criterion of performance is the proportion of errors in the printed copy, and this is determined by the proportion of telegraph elements that are of the wrong kind at the instants of sampling. This error liability cannot readily be deduced from telegraph distortion measurements. In laboratory tests, however, the element error rate can be measured directly by comparing, element by element, the received signals after regeneration with the signals as transmitted, and counting the errors, and the total number of elements, in a test period. Apparatus operating in this way has been in use since March, 1955, and the results have shown the error-counting technique to be a powerful one. It

The authors are at the Post Office Research Station.

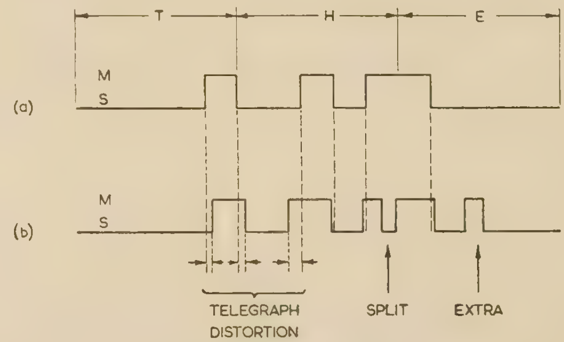


Fig. 1.—Example of radiotelegraph signal (5-unit code).
(*a*) Perfect signal.
(*b*) Disturbed signal from receiver.

has helped towards a better understanding of the noisy-signal performance of radiotelegraph receivers,[1] and has shown that designs based on distortion considerations may not give optimum performance in regenerative systems.

(2) PRELIMINARY CONSIDERATIONS

When the performance of equipment is being measured in the laboratory it is desirable that the conditions of test correspond closely to those encountered in normal use. Thus test equipment for synchronous radiotelegraph systems should cater for operation in the telegraph speed range used in normal service, say 40–180 bauds; in addition, however, it should allow tests to be made at much higher speeds, to cover possible developments and to permit investigation of limiting conditions of operation. Likewise it should cater for a wide variety of test signals, including simulated traffic for normal assessments and repetitive signals of variable mark/space ratio for detailed studies of circuit behaviour. Since radiotelegraph systems usually carry 2-condition, or binary, signals, it is appropriate to design test equipment on a binary-signal basis; 2-channel 4-condition systems,[2] which are not uncommon, can be tested in terms of the individual binary channels after they have been separated.

The special problems of fading, noise and multi-path propagation which arise in long-distance radio links have led to the development of a fading machine[3] to simulate these effects in the laboratory. It is an essential part of the radiotelegraph test equipment, but since its design depends more on considerations of ionospheric propagation than on telegraphy requirements, the improved version in current use has been made the subject of a separate paper.[4]

There are several reasons why element error counting was adopted instead of the obvious method of counting the character errors in printed copy. First, the element-error method is, from the equipment aspect, simpler, and it lends itself to automatic

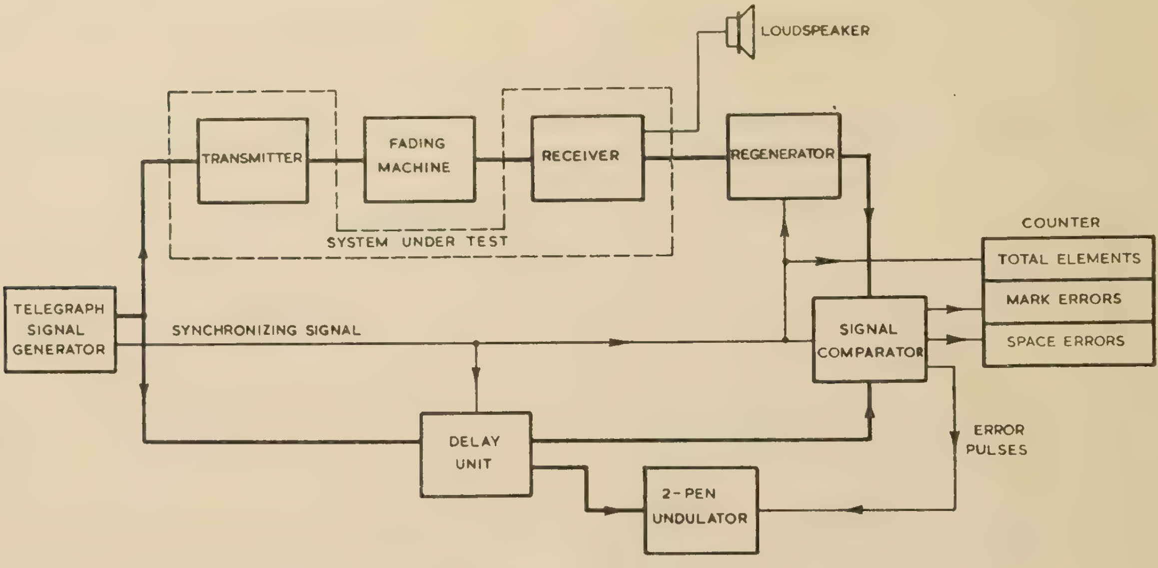


Fig. 2.—Arrangement of test equipment.

counting and to automatic discrimination between mark and space errors. Secondly, it gives more readily the desired flexibility in speed of working. Thirdly, unlike tape readers and machine printers, it can be completely silent in operation. Element error counting does not, of itself, cater for the observation of characteristic errors in mixed signals, or for recording the distribution of errors in time; thus it is desirable that error-counting facilities be supplemented by recording arrangements so that, when necessary, failures can be analysed in detail.

(3) DESCRIPTION OF APPARATUS

(3.1) General

The arrangement of the test apparatus is shown in Fig. 2. Signals from a telegraph-signal generator control a transmitter, the output of which is passed through the fading machine—to simulate propagation—to a receiver. The transmitter and receiver constitute the system under test, and the receiver may well be the complete unit as used at radio stations. Fortunately the high-power stages of radio transmitters usually pass telegraph signals without significant distortion, so that there is no problem of reproducing transmitter distortion in the test equipment; a rudimentary 'transmitter', usually consisting of a low-power transmitter drive unit, suffices. The received signals are regenerated, by inspection at the centre of each element, and are then compared with perfect signals from the telegraph-signal generator. The signals are delayed in passing through the system under test, since the effective bandwidth of telegraph receivers is small, and it is necessary to delay the perfect signals before feeding them to the signal comparator. Whenever an error occurs the signal comparator produces an error pulse. Separate counters are provided for mark errors, i.e. marks received as spaces, and for space errors, and a third counter records the total number of elements in a test period. A 2-pen undulator may be used to record the total error pulses, i.e. mark and space errors together, simultaneously with the perfect signal from the delay unit.

The undulator, which is a standard commercial product, and the counter unit, which uses multi-electrode cold-cathode counter tubes in conventional circuits, call for little comment. The capacities of the error counters and of the total-elements counter are respectively 9 999 and 99 999. The counters are started and reset manually; stopping may be manual, or an automatic stop may be used, functioning after 10^2, 10^3, 10^4 or 10^5 elements. The counters will operate at any speed up to 400 bauds.

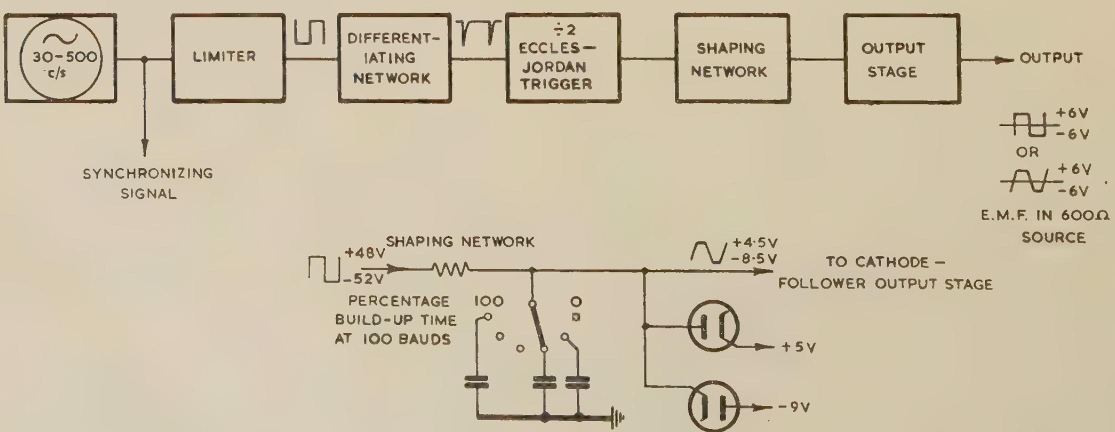


Fig. 3.—Reversals-signal generator.

(3.2) Telegraph-Signal Generators

Three signal generators have been designed for use with the error-counting equipment. One generates reversals, i.e. alternate marks and spaces; another generates signals in cyclically repeated groups of up to seven elements, the numbers of mark and space elements being variable; the third generates a random telegraph signal. All three units deliver a double-current output of trapesoidal waveform with adjustable build-up time; measurements have shown that, from the aspect of economy in bandwidth, this is the optimum modulation waveform for frequency-shift telegraphy.[5]

Fig. 3, which shows the arrangement of the reversals signal generator, is self-explanatory. Signal shaping is obtained by means of an integrating network followed by diode limiters; build-up times of 0, 0·1, 0·2, 0·4, 0·5, 1, 2, 3, 5, 7·5 and 10 millisec are available. A cathode-follower output stage provides an output e.m.f. of ±6 volts in a resistive source impedance of 600 ohms. Another cathode-follower gives a sinusoidal output, from the oscillator, for synchronizing purposes. Ten speeds are provided, in the range 30–500 bauds.

The essential difference between the variable-mark/space-ratio generator and the reversals generator is that the former has a ring counter of cold-cathode valves following the oscillator; the number of valves in the ring is variable over the range 2–7. The firing of one valve in the ring sets the output to 'mark', and any other valve may be selected to restore the output to 'space'. The speed range of the unit is 20–200 bauds in steps of 2 bauds.

In the random-signal generator (Fig. 4) the output of a noise source is sampled by a regenerator at a frequency corresponding to the desired signalling speed, the range of 5–500 bauds being covered in 5-baud steps. Noise generated in a gas-discharge tube is amplified to an effective amplitude of about 35 volts r.m.s., ignoring the effect of peak clipping in the amplifier, in the frequency range 0–250 c/s, before being applied to the regenerator,

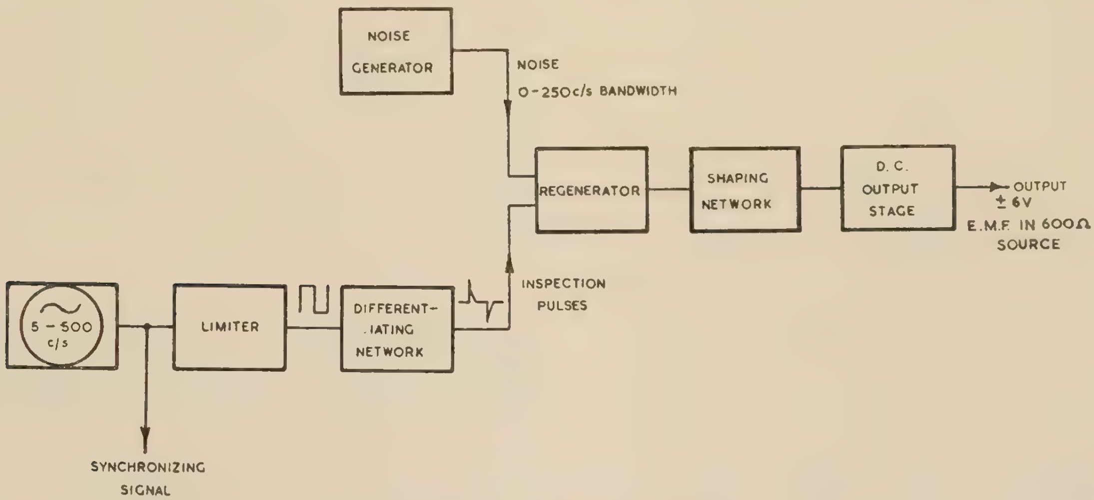


Fig. 4.—Random-signal generator.

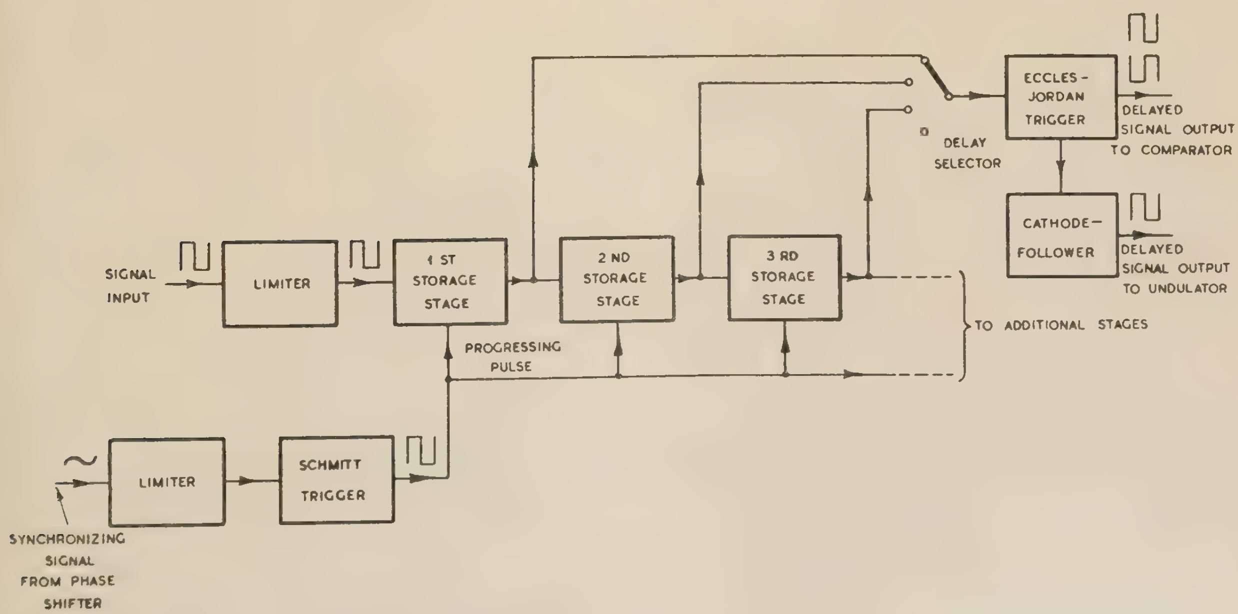


Fig. 5.—Delay unit.

which is similar to the unit described in Section 3.4. A bias control is provided so that the mean frequencies of marks and spaces may be balanced.

In addition to control by their internal oscillators, the signal generators have a facility for control by an external synchronizing signal. This is useful when it is desired to operate at speeds that are not available internally, or when two signal generators have to be run synchronously, as in tests on 2-channel 4-condition synchronous systems.

(3.3) Delay Unit

The delay unit (Fig. 5) incorporates a storage chain of six double triodes, along which the signal is progressed, element by element, by pulses derived from the synchronizing signal. Each stage delays the signal for a time equal to the duration of one signal element, and the total delay is determined by the number of delay stages used; fine adjustments of delay can be made by means of a phase-shifter in the synchronizing-signal input circuit. The delay stages (Fig. 6) are bistable triggers of the Eccles–Jordan type, and the arrangement is such that a stage can change over only when it receives a progressing pulse, and then only if it has been resting in the opposite condition to that of the previous stage. This dependence on the previous stage is secured by d.c. coupling, which increases the trigger sensitivity of the controlled stage if its state differs from that of the controlling stage. Some sluggishness is necessary in the interstage coupling, to prevent a change-over of the controlling stage from causing immediate change-over of the controlled stage; this is obtained by capacitors shunting the anode-grid cross-coupling resistors of the trigger. These capacitors also accelerate a change-over once it has been initiated. The progressing-pulse line, which carries a square wave, is coupled to the anodes, and so to the grids, of the delay stages by small capacitors, and it is the negative pulses produced by the negative-going edges of the progressing square wave that initiate change-over. The immediate effect of a negative progressing pulse is to block the conducting valve of a delay-stage pair, and this causes a positive pulse to appear at its anode. The magnitude of the positive pulse is determined by the supply voltage and by the circuit time-constants and coupling-component values. It is transferred to the grid of the non-conducting valve, where, if the trigger has been sensitized by the previous stage, it will cause the valve to pass current and so initiate a change-over. The condition of the selected delay stage controls an Eccles–Jordan trigger, which provides outputs of normal and reversed sense for feeding to the comparator. Another output, obtained via a cathode-follower, is available for undulator recording. The unit functions at speeds up to 500 bauds.

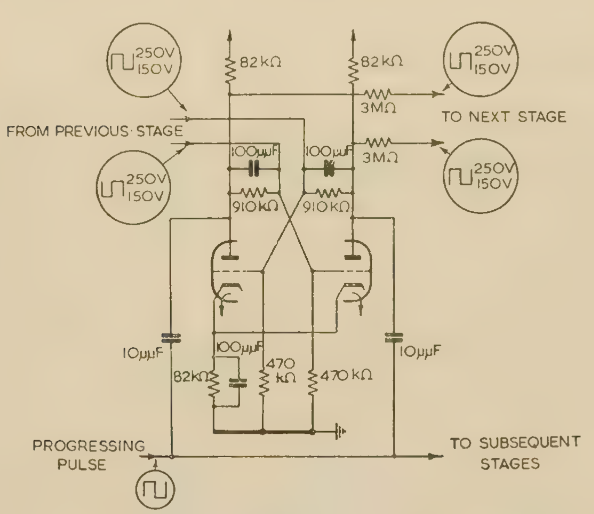


Fig. 6.—Circuit of a delay stage.

(3.4) Regenerator

In the regenerator (Fig. 7), sampling pulses derived by squaring and differentiating the synchronizing signal are applied to the grids of a double triode, V_1, but the common cathode circuit is normally held at +23 volts, so that neither triode conducts. If the signal input is zero, the positive sampling pulses cause both triodes to pass equal currents, and equal negative pulses thus appear in the two anode circuits. If the signal input is other than zero, the balance is upset and unequal pulses are produced in the two anode circuits, the sense of the inequality being determined by the sign of the applied signal. The negative pulses

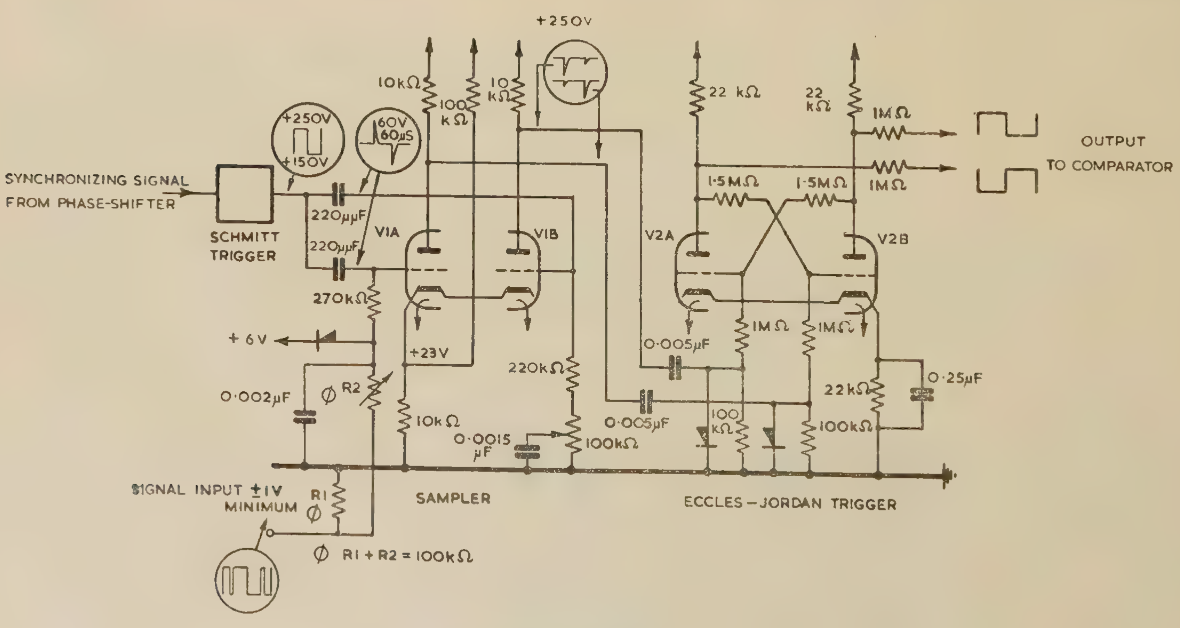


Fig. 7.—Regenerator.

are fed to the two grid circuits of an Eccles–Jordan trigger. Under balanced conditions, i.e. with no signal input, the magnitude of the pulses is sufficient to cut off anode current in both halves of the trigger, and ideally it would be a matter of chance which side recovered first and so took control. A preset control enables a sufficiently accurate balance to be obtained despite asymmetry in the valve or component characteristics. When the negative pulses are unequal the side taking the smaller pulse recovers first, so that the state of the trigger after the sampling depends on the sign of the signal input at the instant of sampling. Outputs are taken from both sides of the trigger, so that signals of normal and reversed sense are available for the comparator.

The rectifiers shunting the 100-kilohm resistors, across which the negative pulses appear, in the trigger grid circuits prevent the accumulation by the $0{\cdot}005\,\mu$F coupling capacitors of charges dependent on the state of the trigger; without the rectifiers there is a memory effect which tends to inhibit change-over.

(3.5) Comparator

The comparator contains two error detectors, one for mark errors, i.e. for mark signals incorrectly interpreted as spaces by the system under test, and one for space errors, each detector comprising a coincidence circuit using three triodes. The grid circuits of two valves are fed with signals, one from the regenerator and the other from the delay unit, one signal being of reversed sense. The third valve is fed with adjustable-phase inspection pulses derived from the synchronizing signal, and the bias conditions are so arranged that this valve can pass current only when a positive inspection pulse occurs while the grids of the first two valves are both in the more negative condition corresponding to a mark or space error. The use of the inspection-pulse technique in this unit avoids the need for accurate phase alignment of the regenerated and delayed signals; it is necessary only that corresponding elements in the two inputs overlap sufficiently to give time for the inspection pulse.

(3.6) Alignment and Monitoring Facilities

Alignment is carried out using a double-beam oscillograph, a display switch being mounted on the equipment to give rapid access to the more important points. First, the phase-shifter associated with the regenerator is adjusted so that the regenerator inspection pulses fall centrally in the elements of the signals received from the equipment under test. Then the delay-unit phase-shifter is adjusted so that the progressing pulses are well clear of the signal transitions, and the number of delay stages is chosen to give an overlap of at least half an element between corresponding elements of the output signals of the delay unit and the regenerator. Finally, the phase-shifter controlling the comparator inspection pulses is so set that the pulses occur well within the overlap periods. Representative waveforms are shown, with a common time scale, in Fig. 8.

Considerable variations can occur in the phasing of the signal from the delay unit, and of the comparator inspection pulse, without any effect on the error counting; the sudden onset of apparent errors makes it obvious when the limit has been reached. The monitor oscillograph is normally kept running continuously when the equipment is in use, and care is taken to keep the regenerator inspection pulses properly phased relative to the incoming signal; in tests involving speed changes, which affect this phasing, frequent adjustment may be necessary.

(4) RESULTS

The error-counting equipment has been used in investigations of several systems, and it has helped towards an improved understanding of the noisy-signal performance of radiotelegraph

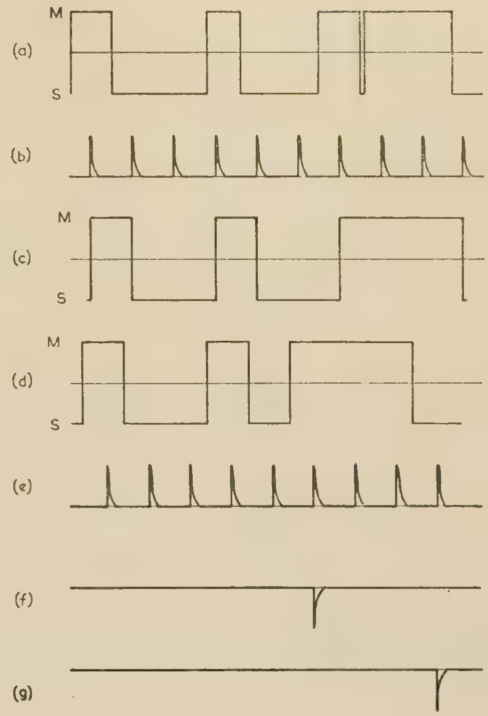


Fig. 8.—Typical waveforms.

(*a*) Disturbed signal from receiver.
(*b*) Regenerator inspection pulses.
(*c*) Regenerator output.
(*d*) Perfect signal, suitably delayed.
(*e*) Comparator inspection pulses.
(*f*) Mark-error pulse.
(*g*) Space-error pulse.

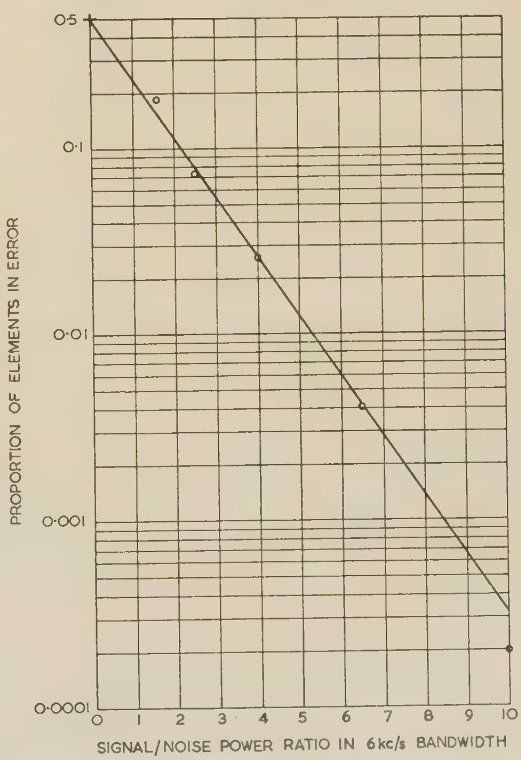


Fig. 9.—Steady-signal performance of on/off receiver on 100-baud reversals.

receivers. Earlier papers[1,6] may be consulted for detailed results, but it is useful to record here some supplementary results obtained since the earlier papers were written.

Recent work suggests that the exponential relation between signal/noise ratio and error rate, that has so simplified the performance rating of frequency-shift-telegraph receivers,[1] may also apply in on/off telegraphy. Fig. 9 shows the results of a steady-signal test of an on/off radiotelegraph receiver, plotted with linear signal/noise and logarithmic error-rate scales. The measured points lie close to a straight line passing through the point (0, 0·5). Further work is needed to establish whether this result has any general validity, and whether its relation to the fading-signal performance is similar to that observed in frequency-shift telegraphy.

The effect of change of telegraph speed on system performance has been discussed in the earlier papers; typical results, obtained with test signals consisting of reversals, show an optimum speed above which the performance deteriorates, owing to the signal elements being shorter than the build-up time of the receiver filters. In practice, the severity of the deterioration at high speeds must obviously depend on the proportion of signal elements that differ from their immediate neighbours, i.e. on the proportion of single isolated mark or space elements in the signal. The reversals signal gives the most severe test. Since the comparatively recent completion of the random-signal generator it has become possible to perform speed-variation tests with random signals, and Fig. 10 shows the results obtained on a demodulation

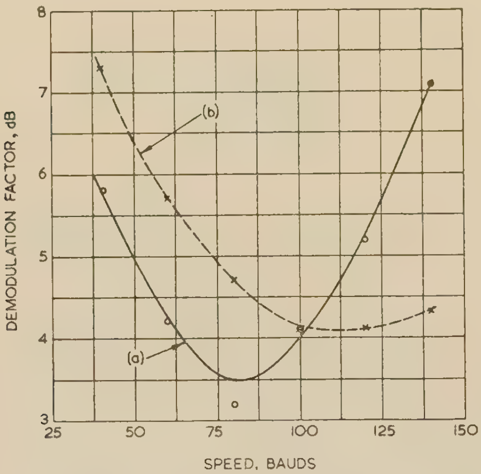


Fig. 10.—Effect of speed variation on performance of experimental receiver.

(*a*) Reversals test signal.
(*b*) Random test signal.

unit forming part of an experimental frequency-diversity receiver;[6] results obtained with a reversals signal are also shown. The performance is plotted in terms of the 'demodulation factor', which is a measure of the amount by which the receiver falls short of the performance of an ideal receiver.[7] It will be seen that with the random signal the optimum speed is higher and less sharply defined.

(5) DISCUSSION AND CONCLUSIONS

The effect on system performance of variation in the proportions of signal elements falling into like-element groups of various lengths shows that, in equipment tests, it is desirable to use signals closely resembling those used in service. The present test equipment leaves something to be desired in this respect. The signals provided by the reversals and variable-mark/space-ratio generators are needed for investigations of circuit details, especially in development work, but they are very different from practical mixed signals. A much better resemblance is obtained with signals from the random-signal generator. Here discrepancies arise from the orderliness of practical signals, owing to language and coding considerations. The 7-unit error-detecting codes[8] provide an example of particular importance. In one code of this type every character is made up of four marks and three spaces, and it follows that like-element groups containing more than eight marks or six spaces cannot occur. Table 1 shows the proportions of elements falling in groups of various numbers of like elements for a completely random signal and for a random message in the 4 : 3 code.

Table 1

PROBABILITIES OF LIKE-ELEMENT GROUPS OF VARIOUS SIZES

Number of elements in group	Proportion of elements in groups of given size		
	Random signal, mark or space	Random 4 : 3 message	
		Mark	Space
	%	%	%
1	12·5	11·6	16·9
2	12·5	16·3	15·9
3	9·4	14·7	7·9
4	6·2	9·1	1·7
5	3·9	3·3	0·5
6	2·3	1·5	0·07
7	1·4	0·5	0
8	0·8	0·09	0
>8	1·0	0	0
	50	57	43

A random-signal generator of the present noise-sampling type could easily be given a preference for the mark condition by appropriate bias at the sampling point. The resemblance to 4 : 3 signals could be improved further by providing a facility within the generator to prevent prolonged holding of the mark or space condition.

Apart possibly from this question of the effective imitation of practical signals, the test equipment satisfies present requirements very adequately. A useful margin of signalling speed is available, and should higher speeds be required in the future, all the hard-valve devices should be adaptable quite easily; the delay unit has, in fact, already been tested at speeds up to 10 000 bauds on reversals. No special virtue is claimed for the circuit techniques, however; they were adopted because, being familiar, they offered the best prospect of producing reliable and effective equipment quickly. Most of the operations performed in the equipment have equivalents in digital computers, and it is likely that a designer experienced in the use of square-loop magnetic materials, for example, could produce a design which was both more compact and more elegant.

In conclusion, it can be claimed that the error-counting test equipment is easy to use, reasonably simple to design and construct and a powerful aid to the system designer. Since the accuracy of the printed copy, by which systems are judged in practical operation, is simply related to the element-error liability, it would be logical to specify the performance of synchronous regenerative radiotelegraph systems in terms of this error liability rather than telegraph distortion as hitherto. Two main advantages are to be expected from such a change: first, the assessment

of equipment performance would immediately become simpler and more realistic; secondly, it should ultimately lead to improved equipment, for design methods based on distortion considerations do not generally lead to optimum performance in respect of error liability.

(6) ACKNOWLEDGMENTS

Acknowledgment is made to the Engineer-in-Chief of the Post Office and to the Controller of H.M. Stationery Office for permission to publish the paper.

(7) REFERENCES

(1) LAW, H. B.: 'The Signal/Noise Performance Rating of Receivers for Long-Distance Synchronous Radiotelegraph Systems using Frequency Modulation' (see page 124).

(2) COLE, A. W.: 'H.F. Communication Systems', *Electronic Engineering*, 1954, **26**, p. 232.

(3) BRAY, W. J., LILLICRAP, H. G., and OWEN, F. C.: 'The Fading Machine and its Use for the Investigation of the Effects of Frequency-Selective Fading', *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 283.

(4) LAW, H. B., LEE, F. J., LOOSER, R. C., and LEVETT, F. A. W. 'An Improved Fading Machine' (see page 117).

(5) ALLNATT, J. W., and JONES, E. D. J.: 'An Investigation of the Spectra of Binary Frequency-Modulated Signals with Various Build-Up Waveforms' (see page 111).

(6) ALLNATT, J. W., JONES, E. D. J., and LAW, H. B.: 'Frequency-Diversity in the Reception of Selectively Fading Binary Frequency-Modulated Signals, with special reference to Long-Distance Radiotelegraphy' (see page 98).

(7) LAW, H. B.: 'The Detectability of Fading Radiotelegraph Signals in Noise' (see page 130).

(8) HAYTON, T., HUGHES, C. J., and SAUNDERS, R. L.: 'Telegraph Codes and Code Convertors', *Proceedings I.E.E.*, Paper No. 1585 R, November, 1953 (**101**, Part III, p. 137).

DISCUSSION ON THE ABOVE PAPERS BEFORE THE RADIO AND TELECOMMUNICATION SECTION, 14TH NOVEMBER, 1956

Mr. A. W. Cole: The papers are indicative of the trend in h.f. communication towards a position where there is little likelihood of major improvements in this method of transmission, and we are now faced with improving the systems by detailed changes in the apparatus, and particularly the receiver.

It is a subject which has been studied for many years, and it is unfortunate that earlier workers have not published their results in the way that the present authors have. The h.f. art has been progressing steadily since its commercial inception in 1926, and even before the war receivers roughly of the type described in these papers were being experimented with. In particular Smale, Copper and Humby* worked on these problems before the war and developed a system known as 'double-frequency keying', where in the receiver-detector advantage was taken of the frequency-diversity characteristics. Receivers of this general type were manufactured, but the technique was dropped and all development since the war has been based on the limiter-discriminator type of receiver, which, as the authors rightly point out, has considerable disadvantages in not making full use of the information contained in the signal.

On the improvement in performance, as I understand it, under the conditions of double diversity, using this type of detector, a gain of 6dB is claimed, which should be divided into its component parts in respect of causes. It must be remembered that this type of detector, as compared with a limiter-discriminator, is essentially a narrow-band device, taking only two small portions of the total band of the receiver. The gain due to this effect is of the order of 3–4dB, and it is the balance of the claimed improvement which is due to the frequency-diversity effect.

Has any consideration been given to the possible effect of very strong adjacent-channel interference? Tests show that, with a limiter-discriminator receiver, this can rise to within 1dB of the wanted signal, and it is still possible—at least under conditions of normal flat fading—to reproduce the telegraph signal, whereas it might be expected that with the authors' type of detection the effects would be much more serious.

The system may prove troublesome in respect of overall bandwidth. The bandwidth of one of the demodulating filters could itself contain an f.m. signal carrying both mark and space frequencies. It is usual to have a trapezoidal waveform for h.f. communication, but we must consider whether we ought not to emulate the practice in the line field and reduce telegraph channels to the minimum bandwidth employing sine-wave modulation.

I am surprised to see that an error rate of 1 in 1000 characters is sufficient to cause slips-twice working on a telegraph circuit. This seems to me to be wide of the mark. On a printer circuit one error every 2mins is handled by the normal RQ manual system. Perhaps the authors are unduly pessimistic.

Diversity switching on limiter-discriminator receivers is usually carried out by the discriminator, but operated by reference back to the amplitude of the signal prior to the limiter, and with certain types of selective fading, with one diversity path for all the mark signals and another for all the space signals, it is over-simplifying the argument to take the single-path case only. Nevertheless, the general trend with this type of detector is beneficial, and we have recently operated it on ionospheric scatter circuits. While we are not able to give full information on the improvement, the use of the mark-and-space detector has reduced the error rate to about 20% of that obtained on the limiter-discriminator type of circuit.

Mr. J. Heaton-Armstrong: The authors of paper No. 2151 (page 98) have compared their receiver, which has a demodulation factor of 3dB, with one having a factor of 9dB. One decibel is accounted for by the superiority of their combining system, but what about the remaining 5dB? Were the bandwidths of the i.f. and low-pass filters in the comparison receiver too large, and will the authors give these bandwidths? The new system would seem to have an advantage during low signal/noise conditions, because the bandwidth of the circuit just before the detectors is narrower and some of the noise products which would fall within the low-pass filter band in the normal system are eliminated.

An assessor circuit is shown in Fig. 6, and the authors mention that it will fail on long mark or space signals. This can easily be avoided by partial d.c. coupling. When a mark and a space assessor are connected together the charges on the condensers will cancel if flat fading occurs, even if it is fast. This is essential for satisfactory operation. I believe that some of the earlier equipment did not have this feature and blocking on flat fading could occur. With the authors' system, however, it is possible for blocking to occur under conditions of fast selective fading

* COPPER, E. G., HUMBY, A. M., and SMALE, J. A.: British Patent No. 480289.

with strong signals; this can produce errors which would not occur with a conventional receiver. The success of the authors' system thus depends on the selective fading being slow compared with the keying speed.

The authors used a filter (Fig. 4) with linear build-up time. How great is the advantage gained by using this filter instead of a pair of critically coupled circuits? Do the authors consider that the performance of their equipment has been degraded by having the limiter after, instead of before, the detector?

Capt. C. F. Booth: The papers clearly demonstrate the soundness of the authors' step-by-step approach. Their reception technique extracts more information from the received signal than the more conventional systems, and, almost equally important, they have outlined a firm basis for comparing the performance of receiving equipments.

I was particularly interested in the improvement of the fading machine since it was described before this Section in 1947. The new version gives a closer approximation in the laboratory to actual conditions on radio circuits, and much of the work described would not have been possible without it. The results of the field tests carried out in 1955 are encouraging, and I wonder whether the authors have information on recent trials. In any case, it is very desirable for future results and achievements to be made public.

In most developments there is something to pay, and here it is in respect of transmitter frequency stability. Section 2.2 of paper No. 2151 (page 98) suggests the mutual frequency difference between transmitter and receiver should not exceed some 25 c/s; for a transmitter working on 15 metres this allows a mutual frequency difference of about 1 part in 10^6. That is achievable, but, as the authors point out, the present international frequency tolerance for the transmitters concerned is 30 parts in 10^6. At the VIIIth Plenary Assembly of the C.C.I.R. in 1956 the revision of this tolerance was considered, and it was decided to recommend within about five years a value of 15 parts in 10^6. This would still not meet the authors' needs, and perhaps they will say how far their present system could be applied to existing transmitters and receivers and the extent of any modification required.

Finally, automatic RQ systems are coming into use on the more difficult circuits, and some may say that this is the answer to the problem of improved performance; it is in part, but we must still improve reception technique to the utmost.

Mr. S. H. Browning: A method of combining the i.f. outputs of several receivers whose aerials are some distance apart has recently been developed. During experiments with the system it has become apparent that the circuits evolved may have application in certain circumstances to diversity reception.

Ordinarily, the outputs of two receivers should not be connected together before demodulation, because the phase relationship is uncorrelated, owing to ionospheric effects and the aerial spacing. However, if some form of phase-correcting circuit is introduced, so that the two signals are in phase, certain advantages are obtained. Using the circuit shown in Fig. A it is possible to add two or more signals before demodulation; the system is an arrangement for transferring the modulation from the received wave to a local carrier, *pt*, and then adding the outputs.

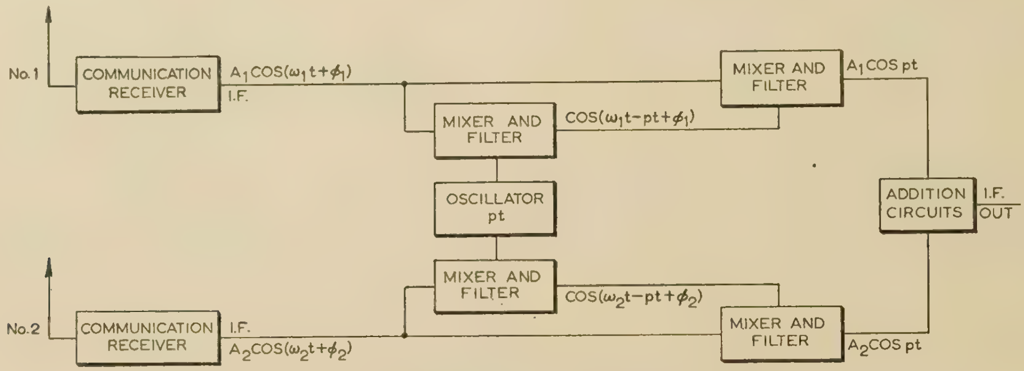


Fig. A.—Diversity system permitting addition before demodulation.

One advantage of addition before demodulation might be an improvement in the signal/noise ratio, since the signals are added in phase whereas the noise powers are uncorrelated. Another feature is that the two signals which are added are at the frequency of the local oscillator. Therefore, circuits of high selectivity can be used at this point which will discriminate against interfering signals at frequencies which differ from that of the oscillator, provided that their amplitude at the input to the combining circuit is smaller than that of the desired signals.

In consequence of these characteristics an improvement of the signal/noise ratio may be possible, precise tuning adjustment of the communication receiver is not necessary, and close control of the i.f. and transmitter frequency will not be required.

Mr. J. D. Holland: Section 3 of Paper No. 2151 (page 98) discusses methods of combination. It is therefore of interest to compare three methods of diversity reception,* namely the selector, linear addition of the contributions, and combination of the outputs in the proportion of the squares of their signal/noise ratio, as shown in Fig. B.

It would appear from Fig. B(i) that, for a signal-level difference greater than 8 dB, the selector method is superior to linear addition; this is an incomplete picture, and Fig. B(ii) shows the comparison based on considering the probabilities of all possible combinations of signal levels at the two receivers. The effect due to diversity reception is seen to shift the distribution laterally into regions of higher signal/noise ratios, the shift being greater for combining systems than for the selector method. The improvement due to ratio squaring is significant, since it is only some 0·7 dB below that which might have been attained with twice the transmitted power.

The error-counting technique given in Paper No. 2259 (page 141) is useful, but I have found another method of examining system design based on the measurement of structural information. In Fig. C waveform (i) is the message and waveform (ii) is the message perturbed by noise, which is fed into a flip-flop circuit arranged to change state close to the zero level line. The

* Johnson, D. A. H.: 'Statistical Approach to the Signal/Noise Improvement with Dual Diversity Receiving Systems', New Zealand P.T.T. Laboratory Report No. 26.

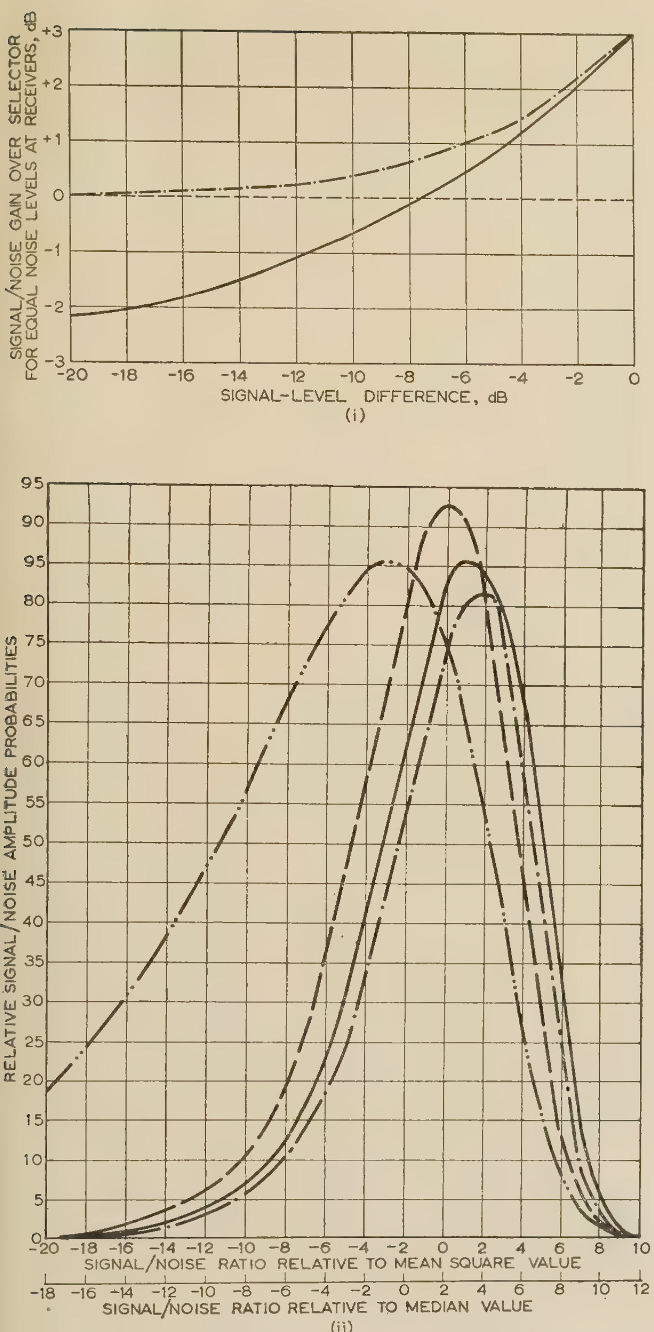


Fig. B.—Comparison of three systems of diversity reception.

——— Linear adder.
— — — Selector.
— · — · — Ratio squarer.
— ·· — ·· — Non-diversity.

(i) Equal signal/noise ratio in both paths.
(ii) Raleigh distribution.

output [waveform (iii)] is differentiated and fed into a counter designed to give a d.c. output proportional to frequency. At 50 bauds the count on an unperturbed message is, on the average, 13 c/s (with a spread of about 11–15 c/s). If the signal/noise ratio is gradually deteriorated, the count reaches an upper limit of about 55 with zero signal input level, which is measured in a bandwidth of 100 c/s between half-power points. Atmospheric noise was used in this test, and it is not clear how the value due to noise alone is obtained.

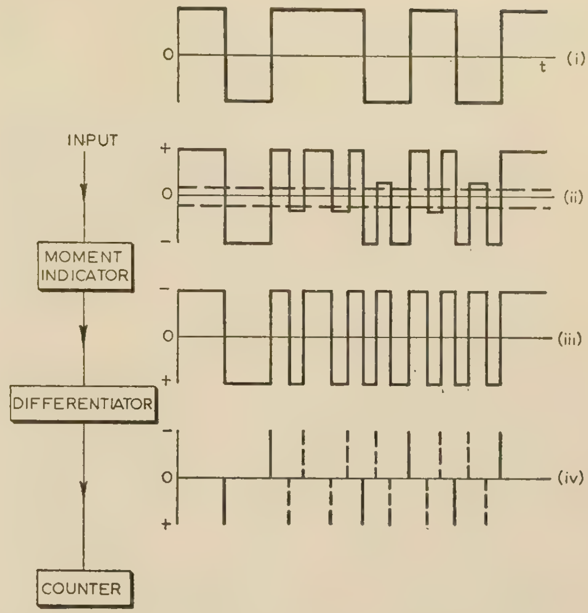


Fig. C.—Measurement of structural information.

Mr. V. J. Terry: It is interesting to note that dual-path transmission is a blessing in disguise. It will make a big difference to radiocommunication if it is now accepted that selectively fading signals can yield more certain results than those subject only to flat fading.

Paper No. 2152 (page 111) bases comparisons on the entire build-up time instead of the customary 19–90%. This seems unfair; the latter may be inappropriate for signals intended for regeneration after reception, but there must be limits of amplitude beyond which further change is unimportant. If the final build-up is an asymptotic approach to the steady-state value, the full build-up time is indefinite and impractical.

The authors' semi-sinusoidal build-up waveform gives sidebands 6 dB worse than the linear build-up of the trapezoidal waveform over an important range of frequencies, but comparison of equal maximum rates of build-up shows that the former is about 2 dB better than the latter. Equal effective build-up times would have given an intermediate result.

Thus there is little to be gained from any reasonable build-up waveform between the linear and the semi-sine-wave; furthermore, if the frequency spread of radio sidebands is to be limited properly, the equivalent of a band filter must be used, and this will inevitably introduce amplitude variation at the transition from mark to space. If the spread is limited at an intermediate frequency, the h.f. stages must have linear amplification to preserve the transient variations of amplitude, the diminution of which will restore the wide spread of sidebands.

The assumption of slow fading is justified by day-time records, but does not the difference between practice and theory arise because the fading is fast when the ordinary receivers are beginning to fail? Moreover, the predominantly north–south transmission used by the authors is probably better from the fading aspect than an east–west one, and the improvement might have been less on the worse routes.

The authors advocate their system for multi-channel working. With eight channels and a transmitter capable of carrying eight frequencies simultaneously without undue intermodulation, it could equally be used for 16 simultaneous frequencies without significant signal-power reduction, thus permitting frequency-shift modulation and dual frequency-diversity for each channel. An f.m. system with 120 c/s channel spacing and 60 c/s frequency

shift working through 100 c/s band-pass filters performs as well as an ordinary 2-tone system with 120 c/s spacing and two 100 c/s filters per channel, and needs no more bandwidth. The multi-channel system will have to be very good to improve on this.

Mr. C. W. Earp: Mr. Browning's diversity system seems to be identical in schematic to that I described in 1951.* In this method, which was designed only for a frequency-shift or other f.m. system, two f.s. signals were converted to corresponding phase-modulated signals at a common output frequency, where they were combined in-phase before limitation and demodulation. However, I fail to understand how the system can provide improved selectivity without corresponding frequency-stability requirements, and I can see no particular advantage of the system when applied to a.m. signals.

I understand from the papers that effective frequency diversity for an f.s. system can be obtained only at the cost of bandwidth corresponding to 500 c/s frequency shift. If such bandwidth is permissible, I should like to mention still another diversity system. Some years ago, having occasion to commutate an h.f. receiver rapidly round a ring of a dozen aerials, I was impressed by the high degree of amplitude diversity among the corresponding 12 signals. The phenomenon suggested that, so long as additional bandwidth for commutation could be permitted, a space-diversity system could be provided as follows. One receiver would rapidly sample the signals from a number of aerials, hence determining which was the strongest, whereby a second receiver could be automatically connected to the 'best' aerial. The system completely avoids the problem of balanced receiver gains inherent in the normal space-diversity system.

Mr. J. A. Smale: Mr. Cole refers to earlier work on this subject. Reports were not published, but our proposals were covered in a British patent.† It is interesting to consider why we failed to produce a system with even that degree of success of the present authors, and it is surprising that nothing has been done in the interim. We were pursuing exactly the same aim, and we failed either because we did not have sufficiently selective filters or—what is more likely—because we were unable to use them since we did not have sufficiently stable transmission frequencies. Perhaps it is not surprising that 20 years ago we did not have a stability of 25 c/s on a circuit working on 15 Mc/s.

A number of interesting points, however, arose from that work, and one of them was that, in failing to produce a frequency-diversity system using the mark and space frequency, we did succeed in producing a telegraph system by combining the mark and space channels in which the final output of the receiver failed in the same direction whether one had extra signals or fades—a feature which facilitated the use of the Verdan system of repetition comprising automatic comparison of two transmissions of each signal element.

The authors quote a maximum multi-path delay of $1\frac{1}{2}$ millisec. This may have been true over the last few years, but it is likely that, in conditions of sunspot maxima, when many more routes are present in the transmission path, there may be much larger path-time delays. In the 1937 tests over the North Atlantic the average was $1\frac{1}{2}$, and the maximum 4, millisec.

H.f. telegraph circuits are rarely if ever dependent on a constant noise level, and it is an important step in the use of the fading machine to introduce recorded noise which is characteristic of such circuits. It is difficult to understand the authors' use of capacitance coupling, because it is dangerous, particularly in the face of strong atmospherics, and there are simple alternatives.

From a telegraphy aspect the pros and cons of placing a telegraph regenerator either at the receiving station or at the telegraph office are very interesting. We considered this point in connection with the provision of the connecting land-line circuits, and it is interesting to note that the performance of these circuits between the Somerton receiving station and the Central Telegraph Office in London is such that the position of the regenerator is unimportant. The bandwidth of the channel between the receiver proper and the telegraph regenerator determines whether the latter should use single-point or multi-point inspection of the signal.

The authors are somewhat severe in criticizing as crude the terms normally used by the maintenance staff to assess the quality of a circuit. I think what they really mean is that there should be a more precise nomenclature for use in laboratory experiments as distinct from those used in everyday working.

Dr. R. L. Smith-Rose: Scientists who were studying the ionosphere many years ago found that they would not get very profitable results by studying fading. They soon found that fading was a very erratic business of a random nature and was generally due to interference between waves reflected from the ionosphere. They then began to study the ionosphere systematically, and the past 25 years have been spent by workers in various parts of the world, pioneered by those in this country, in extracting a systematic knowledge of the ionosphere as a reflecting medium above our earth. It has been implied that this was a blessing in disguise, but some of us have regarded it as a very mixed blessing, and the world might have been a happier place had there been no such thing as the ionosphere; as one result we should probably have had a transatlantic telephone cable many years earlier.

What the ionospheric physicist knows about the ionosphere is that it varies with time and place and season. He understands many of the major variations, but it is unlikely that he will be persuaded to give a generalized answer on the subject of fading. Rather does he say, 'I will tell you what I have discovered and hope that you can make use of it. If you have any questions, please ask them in a systematic and precise form and I will try to give you the answers.'

That being the position, it is clear that any communication engineer who wants to know what fading he is likely to meet must study that fading over the circuits in which he is interested, and that is the particular point on which knowledge must be made available. I have always thought that it was with an appreciation of this limitation that the Post Office research workers many years ago made their fading machine and have used it with the excellent results described in the papers. I hope that they will proceed in that way, gleaning what they can from the ever steadily increasing knowledge of the ionosphere, and applying this knowledge in a direct manner to the particular problems which confront them.

Mr. C. B. Wooster (*communicated*): Paper No. 2103 (page 124) gives the impression that telegraph distortion measurement, both for maintenance and as a means of assessing the performance of a radio link, is superseded by the technique of steady-signal error liability. Pressed to the limit, this could mean a return to the days when the criterion of a circuit in good adjustment was the absence of errors in transmission—at least while the maintenance officer was watching it. Experience has shown, however, that the modern concept of distortion and margin and their practical use in day-to-day maintenance result in incomparably better service. A switching network, such as Telex, would be impossible otherwise.

The laboratory engineer can nevertheless see some objections to it, the most important being the somewhat empirical relationship between measured distortion and the quality of service, and it is here that the measurement of error liability under controlled conditions can be of great value. However, it is necessarily a

* Earp, C. W.: *Proceedings I.E.E.*, 1951, 98, Part III, p. 264.
† Copper, E. G., Humby, A. M., and Smale, J. A.: British Patent No. 480289.

laboratory-type investigation, and at the present stage of the art one may doubt whether the characteristics of a fading machine are acceptable as being in all respects equivalent to actual circuit conditions experienced in practice. Such tests will no doubt go a long way in comparing a new system of transmission with an existing one, and after checking in the field will greatly assist in deciding how much distortion can be tolerated while still providing the necessary grade of service. From this time onward, when presumably the equipment is in service, only distortion measurement can give the maintenance engineer the assurance he needs that the circuit is in good order and will provide good service.

I agree with the author's conclusion that the provision of regenerative repeaters should be regarded as an essential at all radio stations, to correct any distortion present in line tails or terminal equipment, and thus ensures the optimum usage of the radio circuit; distortion measurement will still be necessary, however, to check the performance of the tails and terminal apparatus, and also as a means of assessing the condition of the radio path.

THE AUTHORS' REPLY TO THE ABOVE DISCUSSION

Messrs. J. W. Allnatt, C. G. Hilton, E. D. J. Jones, H. B. Law, F. J. Lee, F. A. W. Levett and **R. C. Looser** (*in reply*): Both Mr. Cole and Mr. Smale refer to a method of frequency diversity patented before the war. This differs from ours in many ways: the judgment level is fixed for example, and the signals are limited before being combined. The avoidance of limiting prior to combination—a feature which interests Mr. Heaton-Armstrong—permits proper weighting of the signals in our receiver.

The importance of frequency economy is appreciated. Although our experiments have been made with a frequency shift of 500 c/s, it does not follow, as Mr. Earp suggests, that useful frequency diversity can be obtained only with such large shifts. Furthermore, there is the possibility of increasing the telegraph speed, so making better use of the bandwidth. The multi-channel system with interleaved frequencies is another possibility. The system is virtually an aggregate of a.m. transmissions, and, as Mr. Terry points out, it could be replaced by a system—also using frequency diversity—of narrow-band f.m. transmissions. Simple theory suggests that there would be little to choose between the two techniques; properly controlled comparative tests seem to be needed.

Mr. Cole and Mr. Heaton-Armstrong discuss the improvement in performance of the new unit over that of the comparison receiver. The latter has a bandwidth of 1 kc/s at intermediate frequency, and a 120 c/s low-pass post-discriminator filter. The narrow-band i.f. filters of the experimental unit give it a 5 dB advantage, which is additional to any frequency-diversity improvement, and also give it good adjacent-channel suppression. The frequency-diversity advantage itself may amount to 6 dB in 2-aerial working at an element error rate of 1 in 10^4.

Mr. Heaton-Armstrong and Mr. Smale object to the a.c. coupling in the assessor—a feature that has not caused any trouble on time-division multiplex systems so far. Replying also to Mr. Terry, we have not been embarrassed by rapid fading on the Australian channels, even at times of fade-in or fade-out. Tests on other routes are needed. We have not tried a pair of critically coupled circuits in place of the linear build-up filters, but the use of a single tuned circuit degrades the performance by about 1 dB.

Capt. Booth asks about further tests. A few field trials were made on the Sydney route towards the end of 1956. On a 50-baud start-stop teleprinter channel the experimental equipment gave a relative performance similar to that shown in Fig. 13 of Paper No. 2151 R (page 98); with white-noise interference the absolute performance was roughly 5 dB inferior to that of an equivalent 5-unit synchronous system. Impulse interference was prevalent during another series of tests using a 93-baud 4:3 error-detecting code; the results given in Fig. 13 (*loc. cit.*) again applied at high error rates, this time in comparison with a receiver of 500 c/s i.f. bandwidth, but at lower error rates the performances of the two receivers were virtually identical.

The application of the new technique to existing systems would usually involve separate a.f.c. on the mark and space channels of the receiver if frequent manual retuning were to be avoided; adequate stabilization of the frequency shift would permit a common a.f.c. system. However, an a.f.c. system is a point of weakness in a receiver; when it is captured, even momentarily, by an interfering signal, a prolonged loss of the wanted signal often occurs. In our view it is better to build adequate inherent stability into transmitters and receivers, and to dispense with a.f.c.

Mr. Holland's comparison of different techniques of dual diversity does not go quite far enough, for the density distributions shown in Fig. B(ii) must be adjusted to have equal areas before they can properly be compared. The improvement given by ideal combination over that given by selection is, in fact, 1·5 dB. We thank him for his method of comparison, however, for it has enabled us to clear up a confusion between the effects of selection alone and of selection plus the exponential non-fading characteristic of the f.m. type of receiver. An Appendix (page 140) has been added to Paper No. 2104 R. The technique of counting transitions seems useful as a quick check on a receiver of known characteristics, but for more general work it suffers from the disadvantage that significant results can only be obtained when the error rate is very high.

Replying to Mr. Smale, the statement about path-time spread in the paper is that it rarely exceeded 2 millisec. The following summary of measurements on facsimile pictures between October, 1953, and July, 1956, may be of interest:

PERCENTAGE OF PICTURES HAVING SPREADS OF PATH-TIME DELAY IN VARIOUS RANGES

Route	Distance	Bearing	Percentage having spreads of				
			0·0–0·5 millisec	0·5–1·0 millisec	1·0–1·5 millisec	1·5–2·0 millisec	2·0–2·5 millisec
	km	deg	%	%	%	%	%
New York	5 600	289	54	21	18	5	2
Melbourne	17 000 / 23 000	75 / 255	10	34	35	15	6
Moscow ..	2 500	64	67	19	12	1	1
Nairobi ..	6 800	137	31	29	28	9	3
Tel-Aviv ..	3 500	113	71	18	11	0	0
Colombo ..	8 800	93	36	28	26	8	2
Amman ..	3 600	112	27	23	26	14	10
Average ..	—	—	35	26	26	9	4

The differences between routes have so far been more marked than the changes on a given route; the measurements are being continued. Multi-path components much lower than −20 dB relative to other components are not likely to be important in practice, and they are not recorded, which might account for the absence of large spreads. Turning to regeneration, the utility of multi-point inspection should depend more on the receiver than on the line bandwidth; it would be of little value with receivers of optimally small bandwidth.

We note Dr. Smith-Rose's remarks with respect, but we persist in our hope that ionospheric physicists will produce information useful to communication engineers and having some general validity. Study of oblique-incidence pulse transmissions, for example, could yield information on the relative amplitudes and path-time delays of multi-path signal components, and on their structures in time and space, their Doppler shifts and their arrival angles. There should eventually be an understanding of the relation between these effects and their ionospheric causes, the irregularities and movements about which much is already known.

The 0–100% definition of build-up time, which was adopted for reasons given in Paper No. 2152 R (page 111), is obviously not universally applicable. We agree with Mr. Terry's conclusions regarding the unimportance of small departures from linear build-up, and the necessity of filtering and linear amplification should more drastic reduction of sidebands be desired.

In reply to Mr. Wooster, error liability was originally suggested as a basis for the specification of system performance, but its use might well be extended to the maintenance of receivers; only simple equipment would be required if repetitive test signals were used.

621.385.032.3

Paper No. 2309 R
Mar. 1957

HIGH-TEMPERATURE PROPERTIES OF TUNGSTEN WHICH INFLUENCE FILAMENT TEMPERATURES, LIVES AND THERMIONIC-EMISSION DENSITIES

By R. N. BLOOMER.

(*The paper was first received* 18*th August, and in revised form* 3*rd December*, 1956.)

SUMMARY

A literature search has shown which data are most reliable for use in calculations of the lives of tungsten filaments run at very high temperatures. Some errors and their possible causes have been found in commonly-used data. Filament lives have been calculated and measured for a variety of electrical supply conditions in a range of high temperatures. The fractional thinning at burn-out can be found from a measurement of the ratio of lives under different electrical supply conditions. Hence, actual lives can be foretold accurately from measurements of heating current and initial wire diameter. Thermionic-emission measurements have confirmed that temperatures can be found from heating currents, even for short filaments where correction has to be made for the heat lost by conduction to the supports.

(1) INTRODUCTION

Pure tungsten is still widely used as a thermionic emitter when very high emission current densities are needed, even though many other complex cathodes with lower work functions have been developed. Emission densities of several amperes per square centimetre can be drawn continuously, and in somewhat inferior vacua, since no poisoning effects occur at high temperatures (say, 2 700° K and higher). But since the evaporation of tungsten atoms, like that of electrons, increases exponentially with temperature, large emission densities can only be obtained for short times until the filaments burn out. In other words, such filaments have short lives. (In poor vacua the thinning of filaments may be hastened by gas erosion. This independent cause of short lives will not be treated in the paper.) The paper is concerned with the prediction of filament lives that are limited by evaporation and of thermionic emission densities, particularly for temperatures above about 2 700° K.

The lives of filaments can be calculated if the temperatures of the hottest parts, and the corresponding evaporation rates, are known at all times. In addition, the percentage thinning at burn-out must be known. The method of finding temperatures from heating currents and diameters is well known and accurate.[7] For a filament which is so long that no heat is lost by conduction from the hottest part to the supports at the ends the method and data can be used directly, as in the paper. A novel method of finding the percentage thinning at burn-out is described. It is found experimentally that filament lives can be foretold accurately only if some commonly accepted values for the evaporation rate of tungsten are rejected in favour of others. Adventitious support for this rejection has come through a literature search for the origins of the evaporation data.

Thermionic emission is very dependent upon temperature, and at very high temperatures it is not readily influenced by any other variable such as vacuum conditions. Thus temperatures of very hot filaments can be found accurately, even in demountable vacuum apparatus, from measurements of thermionic emission densities. Those measurements give an independent check upon the method of calculating filament temperatures from heating currents and diameters, which is useful for thin wires like those studied in the present work. For thicker wires, diameters can be measured directly, and their temperatures can then be deduced from values of thermionic emission.

(2) THE TEMPERATURE OF TUNGSTEN FILAMENTS

Many workers have found that the temperature of tungsten filaments of known diameters can be conveniently and accurately found by measuring the heating current. Thus both Dushman *et al.*[3] and Reimann[12] have used this means of finding and controlling the temperature in measurements of the thermionic emission from tungsten.

The relation between heating current and diameter can be shown simply, for parts of a wire so remote from the ends that no heat is lost by conduction, by equating the Joule heating power to the power lost by radiation.

For a unit length of wire

$$I^2 \frac{4}{\pi d^2} f_1(T) = \pi d f_2(T)$$

whence
$$I = d^{3/2} \frac{\pi}{2} [f_2(T)/f_1(T)]^{1/2} \quad . \quad . \quad . \quad . \quad (1)$$

where I = Heating current in a wire of diameter d.
T = Temperature.
ρ = Resistivity = $f_1(T)$.
P = Power radiated by unit area = $f_2(T)$.

Since it has been shown experimentally, e.g. in Reference 12, that the temperature of filaments can be found accurately from $I/d^{3/2}$ it is clear that the ratio P/ρ has a reproducible value at any particular temperature. In contrast, it is well known that ρ alone varies with time, e.g. because of combination with carbon in continuously-pumped metal vacuum systems. Thus for reproducibility it is better to base temperature on current measurements, i.e. on the ratio P/ρ, than upon a single specific property like resistivity or spectral emissivity.

Jones and Langmuir[7] have given the corresponding values of $I/d^{3/2}$ and temperature for tungsten wires. Reimann[12] determined the relation between I and T for a particular-diameter wire (0·008 in). The differences between the temperatures deduced by these different workers are small.

Corresponding temperatures are as follows: Jones and Langmuir, 2 000° K, 2 200° K, 2 400° K, 2 600° K, 2 800° K, 3 000° K; Reimann, 2 000° K, 2 205° K, 2 410° K, 2 614° K, 2 819° K, 3 024° K.

The practical significance of the differences is small, as is shown by considering a particular case. For a 0·005 in-diameter wire, heating currents in the range 2·2–3·0 amp give temperatures in the range 2 500–3 000° K. In routine measurements, it is unlikely that the current can be set or read to ±0·02 amp, and this is equivalent to an error of ±12° K. Hence, it does not matter, in practice, which set of current/temperature data is used. Here, the Jones and Langmuir[7] values have been used, since they are well known and widely used by others.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Bloomer is with Associated Electrical Industries Ltd.

(3) THE RATE OF EVAPORATION OF TUNGSTEN

There is a wide scatter in the rates of evaporation of tungsten reported by different workers, as shown in Table 1.

Table 1

A Selection of Published Evaporation Rates at Very High Temperatures

T	Evaporation rate			
	Zwikker[15]	Jones and Langmuir[7]*	Wahlin and Whitney[14]	Reimann[12]
°K	g-cm^{-2} sec^{-1}	g-cm^{-2} sec^{-1}	g-cm^{-2} sec^{-1}	g-cm^{-2} sec^{-1}
2 800	$8{\cdot}33 \times 10^{-8}$	$1{\cdot}12 \times 10^{-7}$	$5{\cdot}9 \times 10^{-8}$	$7{\cdot}20 \times 10^{-8}$
2 900	$3{\cdot}09 \times 10^{-7}$	$3{\cdot}45 \times 10^{-7}$	$2{\cdot}0 \times 10^{-7}$	$2{\cdot}36 \times 10^{-7}$
3 000	$1{\cdot}05 \times 10^{-6}$	$9{\cdot}69 \times 10^{-7}$	$6{\cdot}4 \times 10^{-7}$	$7{\cdot}15 \times 10^{-7}$

* The values in this column have been obtained by dividing the Jones and Langmuir figures by π.

The data of Reimann[12] have been used for the calculations. These then agree with the measurements of filament lives given in Section 6. The difference between the Wahlin and Whitney[14] and the Reimann[12] values is equivalent to a temperature change of only 15° K at 3 000° K; but the difference between the Reimann[12] and the Zwikker[15] values is equivalent to a temperature change of 40° K at 3 000° K, and a consequent difference of 40% in filament lives, on the Reimann temperature scale. Some explanation must be given for the rejection of the Jones and Langmuir[7] values, which are still used by others.

First, the evaporation rates presented by Jones and Langmuir are not directly experimental. They were derived in the following way. Langmuir[9] in 1913 measured the rate of evaporation in the range 2 440–2 930° K, although the most accurate measurements were in the range 2 825–2 930° K. These temperatures were obtained from a comparison of the light emitted by the tungsten with that of a standard lamp, viewed through a special blue glass, by using the formula

$$T = \frac{11\,230}{7{\cdot}029 - \log H}$$

in which H is the intrinsic brilliancy of the tungsten filament in international candle-power per square-centimetre (projected area). This formula was obtained from one given by Nernst[10] in 1906 for a black body, by using the formula $K = 0{\cdot}021\,8H$ to convert from the brilliancy, K, in Hefner candles per square-millimetre to that, H, in international candles per square-centimetre. The Nernst formula had itself been obtained by using experimentally measured data for black bodies, in a theoretical formula derived by Rasch[11] in 1904. Hence there is a possibility of inherent error in the Langmuir[9] formula because of error in the formula for conversion of photometry units. For example, in 1925 Zwikker[16, 17] reported that 1 international candle was taken as equivalent to 1·111 Hefner candles before 1914; but that after the 1914–18 War 1 international candle was taken as being equivalent to 1·15 Hefner candles. It was not known whether either or both units had changed in size. By the time of the publication of the Jones and Langmuir paper[7] in 1927 it was clear that the earlier evaporation rates disagreed with the then recently measured values, e.g. those of Zwikker.[15] In order to bring the earlier evaporation rates into better agreement with the more recent values, big changes were made in the temperature scale. These were reported by Jones, Langmuir and Mackay,[8] and in outline were as shown below:

Temperatures on the Langmuir[9] candle-power scale	2 440° K	2 738° K	2 930° K
Temperatures corrected to agree with Zwikker[15] spectral emissivity scale	2 518° K	2 852° K	3 066° K

Thus the original more accurate measurements of evaporation rate were made in the temperature range 2 950–3 070° K on the new scale. In presenting this modified experimental data, Jones and Langmuir[7] extrapolated over the temperature range 1 000–3 655° K.

Secondly, the manner of presentation of the evaporation rates in Table 1 col. 11 of Reference 7 could be misleading, if the Table were used apart from the text. This makes possible a curious explanation of the still persisting use of their values. The unit given above the column of numerical values is grammes × (centimetres)$^{-2}$ × (seconds)$^{-1}$ and all the values in Table 1 col. 11 of Reference 7 should be divided by π in order to express them in this the conventional unit. It is true that in many subsequent presentations of the evaporation-rate data by others,[4, 6, 13] the Jones and Langmuir[7] figures have been divided by π.

(4) AN OUTLINE OF THE CALCULATIONS OF FILAMENT LIVES

(4.1) Filaments supplied at Constant Currents

For filaments supplied at constant currents the (current/diameter$^{3/2}$)/temperature data of Jones and Langmuir[7] can be applied directly. The temperature corresponding to a particular current is found for the initial diameter of the wire. The rate of evaporation for this temperature is known from the data of Reimann,[12] and hence the time taken for the wire diameter to decrease slightly is calculated. (In the present calculations a thinning by 2% of the initial diameter has been taken as the size of step.) This process of calculation is repeated step by step for a total thinning of the wire by 12%. When the filament is supplied at constant current the temperature increases as the wire thins, and the times for the evaporation of successive layers of equal thickness decrease rapidly. Thus the majority of the lifetime corresponds with the first 4% of thinning.

(4.2) Filaments supplied at Constant Voltages

In order to be able to use the (current/diameter$^{3/2}$)/temperature data of Jones and Langmuir[7] when a filament is supplied at constant voltage, it is necessary to know the ratio of resistivity to its reciprocal rate of temperature variation. It has already been emphasized that the resistivity of tungsten alone is not a reliable basis for the scale of temperature, but the present calculations, as the following outline shows, are only partially dependent upon the ratio $(1/\rho)d\rho/dT$, rather than the resistivity directly. Moreover, at very high temperatures, where lives are limited by evaporation to relatively short times, few impurities will persist in the tungsten.[1]

For a constant voltage V across a filament, of length l, initial diameter d_i and final diameter d_f, the initial and final currents, I_i and I_f, are given by

$$I_i = V\pi d_i^2/4\rho_i l$$

and

$$I_f = V\pi d_f^2/4\rho_f l$$

where ρ_i and ρ_f are the initial and final resistivities. When the wire thins by evaporation the current and temperature will decrease. (The change in length of the filament is negligible.) The initial and final temperatures, T_i and T_f, can be found from the values of $I_i' = I_i/d_i^{3/2}$ and $I_f' = I_f/d_f^{3/2}$. In the small range

of temperature drop during each step of thinning, it is sufficient to assume that I' is linearly dependent upon T.

Thus $$I_i' - I_f' = a(T_i - T_f)$$

Similarly, the resistivity can be assumed to be a linear function of temperature within a small interval,

and so $$\rho_i - \rho_f = b(T_i - T_f)$$

From these and the two previous pairs of equations it can be shown that

$$T_i - T_f = g/(a/I' + b/\rho_i) + g^2(ab/I_i'\rho_i)/(a/I_i' + b/\rho_i)^3 + 2g^3(ab/I_i'\rho_i)^2/(a/I_i' + b/\rho_i)^5 + \ldots \quad (2)$$

in which $g = 1 - (d_f/d_i)^{1/2}$.

In applying these calculations it is supposed that the voltage across the filament is that corresponding to some particular initial current I_i through the wire. T_i is then found from $I_i/d_i^{3/2}$, using the Jones and Langmuir[7] data. The values of ρ_i, a and b are next found, again using data in the Jones and Langmuir[7] paper. The value of $T_i - T_f$ is now found from eqn. (2), using the value of g appropriate to the initial and final diameters in this step of the thinning. The value of I' corresponding to the final temperature is now taken as the value I_i' for the next step of thinning. By repetition, the temperature at the beginning of each step of thinning can be found. Finally, the time for each successive step of thinning is calculated, using the evaporation rates given by Reimann.[12]

A typical example of such calculations is given. For a wire of initial diameter 0·005 in supplied at a constant voltage such that the initial current is 3·0 amp, we have $I_i' = 2096$ amp-cm$^{-3/2}$, $T_i = 3030°$K and $\rho_i = 92 \cdot 8 \times 10^{-6}$ ohm-cm. In the region of 3000°K, $a = 1 \cdot 13$ and $b = 3 \cdot 7 \times 10^{-8}$. For the first 2% of thinning $g = 1 \cdot 01 \times 10^{-2}$. Hence, from eqn. (2),

$$T_i - T_f = 10 \cdot 86 + 0 \cdot 027 + \ldots = 10 \cdot 89° \text{K}$$

Similarly, for the next 2% of thinning, it is found that $T_i - T_f = 11 \cdot 23°$K. For these first two steps the evaporation times are 0·78 and 0·91 hour respectively.

The above calculations have been carried out in small steps because the ratio of the life at constant voltage to that at constant current, as well as the actual lives, increases with the total amount of thinning that occurs before burn-out. Hence, from a comparison of measured and calculated ratios it is possible to deduce the thinning that takes place before the filaments burn out. The conventional lifetime of a filament ends after the initial diameter has decreased by 10% (see Reference 2). If it is only required to find this conventional life, the calculation can be done sufficiently accurately in one step. Thus, for 10% thinning of a 0·005 in wire, initially at a temperature of 3020°K, when supplied at constant voltage, the decrease of temperature calculated by the 2% step process is 56·7°K, whereas if the calculation is made in one step of 10% the temperature drop is 55·8°K. Hence, in practice, the simpler process may be used. Moreover, in calculating the variation of life with the amount of thinning at burn-out, it is sufficiently accurate to suppose that the temperature of the thinning wire falls uniformly with diameter.

(4.3) Filaments Supplied at Simultaneously Varying Voltages and Currents

In many practical cases the type of electrical supply is intermediate between the extreme cases treated in Sections 4.1 and 4.2. Thus, when the sum of the internal resistance of the generator plus the resistance of the connections is n times ($n \simeq 1$) the resistance of the filament supplied, both the voltage across, and the current through, the filament will change as it thins.

In this case it can be deduced that

$$T_i - T_f = x/y + abx^2/I_i'\rho_i y^3 + 4a^2b^2x^3/I_i'^2\rho_i^2 y^5 + \ldots \quad (3)$$

where $$x = nG^2 - nG^{1/2} + 1 - G^{1/2}$$

$$y = a/I_i' + b/\rho_i + naG^2/I_i'$$

and $$G = d_f/d_i$$

As with eqn. (2), only the first term of the expansion (3) need be enumerated, and the calculation of the temperature drop can be made in one step.

For the particular case $n = 1/3$, $T_i - T_f = 0$ for small changes in filament diameter, and so the temperature of thinning filaments remains constant. The temperature of a thinning filament increases when $n > 1/3$, and decreases when $n < 1/3$.

(5) THE RESULTS OF SOME CALCULATIONS OF FILAMENT LIVES

All the calculations are for wire of 0·005 0 in initial diameter.

Table 2

LIVES AT CONSTANT CURRENT

Percentage thinning at burn-out	Lives			
	2·70 amp	2·80 amp	2·90 amp	3·00 amp
%	h	h	h	h
2	6·14	2·90	1·51	0·78
4	9·55	4·56	2·38	1·20
6	11·39	5·49	2·83	1·43
8	12·35	5·99	3·08	1·56
10	12·77	6·24	3·22	1·63
12	13·05	6·38	3·30	1·67

Table 3

LIVES AT CONSTANT VOLTAGE

Percentage thinning at burn-out	Lives	
	$I_i = 2 \cdot 70$ amp	$I_i = 3 \cdot 00$ amp
%	h	h
2	6·14	0·78
4	13·09	1·69
6	20·92	2·69
8	29·65	3·80
10	39·38	5·08
12	50·02	6·58

Table 4

LIVES WHEN CURRENT AND VOLTAGE CHANGE

Percentage thinning at burn-out	Lives		
	$n = 1/3$	$n = 1$	$n = 3$
%	h	h	h
2	0·78	0·78	0·78
4	1·56	1·44	1·32
6	2·34	1·99	1·70
8	3·12	2·48	1·96
10	3·90	2·89	2·15
12	4·68	3·24	2·30

For an *initial* current of 3·0 amp, n = (supply-source resistance)/(filament resistance). For $n = 0$, see Table 3; for $n = \infty$, see Table 2.

From Tables 2 and 3, it is seen that the ratio of the lives at constant voltage and at constant current, for any particular thinning, is almost independent of the initial current through the wire. Hence lives at constant voltage for initial currents other than those given in Table 3 can be calculated readily.

(6) SOME MEASUREMENTS OF FILAMENT LIVES

(6.1) Experimental Method

Experimental valves were prepared containing tungsten filaments 3–5 cm long made from nominally 0·005 in-diameter wire welded to 0·040 in-diameter nickel supports. The tungsten wire was used as supplied by the makers. It is important to check its initial diameter, since the evaporation rate for a given heating current increases exponentially with $(\text{diameter})^{3/2}$. As a rough check, sectioned wire was examined with a projection microscope. As a more critical test, the thermionic emission density for particular heating currents was measured, as described later in Section 7. From these checks, and from measurements of filament lives, it was found that the initial diameter of one batch of wire was 0·0050 in, and of another 0·0048 in.

The experimental valves were exhausted on a glass vacuum system and baked at 400–450° C for a few hours. The pressures within them, measured by an ionization gauge, were less than 1×10^{-7} mm Hg after the bake, and did not increase above 1×10^{-6} mm Hg at any time during the life tests, which were all made with the valves still on the pumping system. At such pressures, the rate of thinning of the filaments through gas erosion (by oxygen and water vapour principally) is negligible compared with the rate set by the thermal evaporation of tungsten. This fact was confirmed by the observation that lives of filaments run at the same current were just as long in runs made on an unbaked demountable vacuum system in which the pressure, of water vapour principally, was about 1×10^{-5} mm Hg.

The filaments were heated by 50 c/s alternating currents from a voltage-stabilized transformer with a 20-volt 20 amp secondary winding, run at only a fraction of its maximum rating. Care was taken to check the truth of current measurements, since, for a 0·005 in-diameter wire in which the heating current is 3·0 amp, an error of 1% in the setting or reading of the current causes a 0·6% error in the wire temperature, for which the equivalent change in evaporation rate, and hence filament life, is 24%.

The total running time before burn-out (i.e. the life) has been measured for filaments whose heating current was kept constant from within a few seconds of first switching on the supply. In the first two or three minutes of running, it was necessary to increase the voltage applied by about 5% in order to maintain a constant current flow, because of crystalline changes which occur at the first heating. Thereafter, in all life tests, the voltage needed to maintain a constant current increased at a much slower rate than that set by the thinning of the wire through evaporation of tungsten. Therefore, in further experiments to find the lives of filaments run at constant voltage, each filament was allowed to run for a minute or two before the voltage required at some selected heating current was measured. Then, for the remainder of the life of each filament, the voltage was kept constant at the selected value. In all the runs, both at constant current and at constant voltage, it was usual to keep the filaments glowing continuously until they burnt out. However, in some cases the runs were interrupted, either for about an hour or overnight. In neither variety of interrupted run was the total running time before burn-out different from that found for filaments run continuously, under the same electrical supply conditions.

(6.2) Results

Figs. 1 and 2 show the results of some measurements. Fig. 1 shows the calculated values of the ratio of lives at constant voltage to constant current as a function of the amount of thinning that occurs before burn-out. The measured value of this ratio was found to be 1·8(1) (the mean, for six pairs of filaments, lying between 1·75 and 2·05). From this and Fig. 1 it is deduced that the filaments burn out when their initial diameter has been reduced by 6% (cf. References 2 and 12). Fig. 2 shows that the agreement between the calculated and measured lives is good.

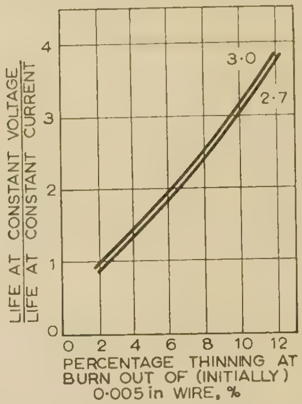


Fig. 1.—Variation of the ratio of the life at constant voltage to that for a constant current supply for various percentage thinnings at the end of life.

Numbers against curves are the initial currents, in amperes, through wire.

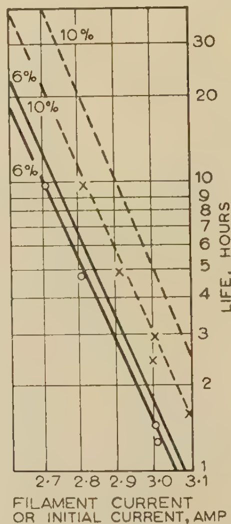


Fig. 2.—Lives of long 0·005 in-diameter filaments for constant-current and constant-voltage operation.

---- Constant voltage } Calculated.
—— Constant current }

The numbers against curves are percentage thinning at burn-out.

× Constant voltage } Experimental points.
○ Constant current }

(7) SOME MEASUREMENTS OF THERMIONIC EMISSION

(7.1) Experimental Method

The thermionic emission from samples of the wire used in the life measurements has been measured in the cylindrical test diode designed and used by Haine and Einstein.[5] The diode was pumped continuously on a demountable vacuum plant. The filaments, of 0·005 in and 0·0048 in diameter, were 5 cm

long and their temperatures over the middle region were found from the heating currents and diameters. Alternating and direct currents were used for heating. Provided that the r.m.s. values of the alternating currents were known truly, the same emission was obtained with equal currents of the two types, and so, in the majority of the work, alternating current was used for heating the filament. An alternating 50 c/s voltage was applied between the three anodes (a central one and two guard anodes, each 1 cm long) and the electrical centre of the filament. The emission current to the central anode, during the positive half-cycle of the applied voltage, was calculated from measurements made with a calibrated cathode-ray oscilloscope of the voltage developed across a known resistance.

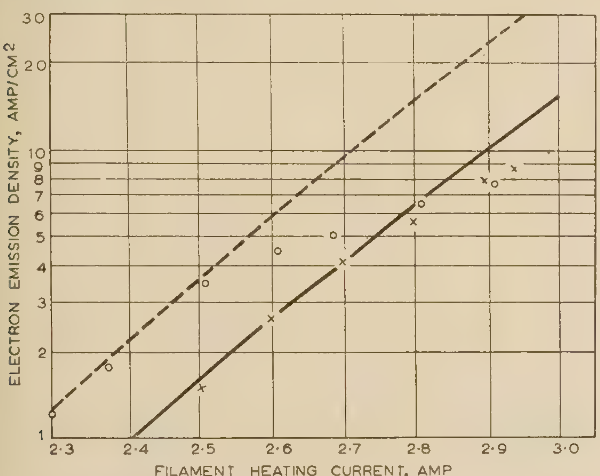


Fig. 3.—Electron emission for various heating currents predicted by Jones and Langmuir.

The observations on test diodes are indicated by crosses and circles.
——— 0·005 0 in-diameter wire.
– – – – 0·004 8 in-diameter wire.

(7.2) Results

Fig. 3 shows the results obtained in typical runs. The emission densities agree with the values computed from the Jones and Langmuir[7] data,* up to a temperature of about 2 800° K, for both types of wire (0·005 0 in untreated; 0·004 8 in cleaned and annealed). At temperatures higher than about 2 800° K the cooling effect of the electron emission was so great that measured emissions were less than computed values by amounts proportional to the computed emission current. The temperature at which the measured electron emission began to be less than the accepted values was lower in the case of the thinner wire (2 800° K for 0·004 8 in; 2 830° K for 0·005 in). This is because the thermal capacity and conduction are smaller for the thinner wire.

(8) CONCLUSIONS

The lives of tungsten filaments can be foretold accurately by using the Jones and Langmuir[7] data for temperature in terms of heating current and the Reimann values[12] of evaporation rate, and by assuming that lives end when the initial diameter of filaments has been reduced by 6%. Lives calculated by using the still frequently accepted evaporation data of Jones and Langmuir[7] are several times shorter than observed, even when using the conventional criterion that life ends after 10% thinning.[2] The initial diameters of filaments can be found accurately both from measurements of the density of thermionic emission at known heating currents, provided that the emission density is not so great as to cause appreciable cooling of the wire, as well as by comparing measured and calculated lives.

It must be emphasized that a reliable scale of temperature can be made in terms of filament heating current only if the initial diameter is known, and if care is taken to make accurate measurements of current. Furthermore, although the calculations of lives can be made for lower temperatures, the measurements were not extended to temperatures below about 2 750° K. At a sufficiently low temperature the good agreement between the measurements and calculations might cease. Not only might gas attack upon the filament become appreciable—although improvement in working vacua would overcome this—but long-term change in the properties of the tungsten might alter the length of life for any given electrical supply condition.

* Thermionic emission densities are given in Table 1 col. 10 of the paper by Jones and Langmuir in the unit ampere/cm² × π.

(9) ACKNOWLEDGMENTS

The author would like to thank Mr. M. E. Haine for helpful discussions, and Dr. T. E. Allibone, Director of the Laboratory, for permission to publish the paper.

(10) REFERENCES

(1) De Vos, J. C.: *Physica*, 1954, **20**, p. 690.
(2) Dushman, S.: *General Electric Review*, 1915, **18**, p. 156.
(3) Dushman, S., Rowe, H. N., Ewald, J., and Kidner, C. A.: *Physical Review*, 1925, **25**, p. 338.
(4) Espe, W., and Knoll, M.: 'Werkstoffkunde der Hochvakuumtechnik' (Springer, 1936).
(5) Haine, M. E., and Einstein, P. A.: *British Journal of Applied Physics*, 1952, **3**, p. 40.
(6) Kohl, W. H.: 'Materials Technology for Electron Tubes' (Reinhold, 1951).
(7) Jones, H. A., and Langmuir, I.: *General Electric Review*, 1927, **30**, p. 310.
(8) Jones, H. A., Langmuir, I., and Mackay, G. M. J.: *Physical Review*, 1927, **30**, p. 201.
(9) Langmuir, I.: *ibid.*, 1913, **2**, p. 329.
(10) Nernst, W.: *Physikalische Zeitschrift*, 1906, **7**, p. 380.
(11) Rasch, E.: *Annalen der Physik*, 1904, **14**, p. 193.
(12) Reimann, A. L.: *Philosophical Magazine*, 1938, **25**, p. 834.
(13) Smithels, C. J.: 'Tungsten' (Chapman and Hall, 1952).
(14) Wahlin, H. B., and Whitney, L. V.: *Physical Review*, 1936, **50**, p. 735.
(15) Zwikker, C.: *Archives néerlandaises des sciences exactes et naturelles* (Series IIIA), 1925, **9**, p. 207.
(16) Zwikker, C.: *Konilijke Akademie van Wetenschappen, Amsterdam*, 1924, **28**, p. 499.
(17) Zwikker, C.: *Physica*, 1925, **5**, pp. 249 and 319.

621.395.625.3

Paper No. 2296 R
Mar. 1957

A SURVEY OF FACTORS LIMITING THE PERFORMANCE OF MAGNETIC RECORDING SYSTEMS

By E. D. DANIEL, M.A., P. E. AXON, O.B.E., M.Sc., Ph.D., Associate Members, and W. T. FROST, Graduate.

(*The paper was first received* 1*st June, and in revised form* 2*nd October*, 1956.)

SUMMARY

Various elements of a magnetic recording system, such as the heads, the tape and the tape transport mechanism, cause departures from the 'ideal' performance of the system, due either to the physical properties of the materials of which they are composed or to the limitations of the accuracy to which they can be made. Some of the effects depend fundamentally only on signal frequency and others only on recorded wavelength. The paper examines the nature and magnitude of the various departures and discusses the improved properties required in the various elements if the ideal performance is to be more closely approached.

(1) INTRODUCTION

In a recent paper[1] Selsted and Snyder discussed some of the limitations of existing magnetic recording techniques and the influence of the magnetic medium upon them. It is the purpose of the present paper to examine elements such as heads, recording media and tape transport systems, which govern the performance of a recorder, rather more closely and so to indicate the difficulties and requirements in various types of application. Fairly detailed treatment is given of some factors which have, hitherto, not been widely discussed. Various factors will be considered in the context of the recording and reproducing processes and with reference to frequency effects (i.e. those depending fundamentally only on signal frequency) and wavelength effects (i.e. those depending fundamentally on recorded wavelength). Frequency and wavelength effects may exist independently or together, depending on the frequency range and recording speed in any particular application. It should be assumed that the paper relates to a tape system, unless otherwise stated, although most of the effects described will also be present, in a more or less exaggerated form, when recording on magnetic drums or discs.

(2) FUNDAMENTAL PROPERTIES OF MAGNETIC RECORDING

(2.1) The Idealized System

A diagrammatic representation of a tape recording and reproducing system is given in Fig. 1A, and a more detailed view of a head is given in Fig. 1B. Let the recording head be fed with a signal current I varying sinusoidally with time at a frequency f. If the core is of infinite permeability, the peak value of the field strength created within the gap is given by

$$\hat{H}_x = (4\pi N'/b')\hat{I} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where b' is the length of the gap and N' is the number of turns on the coil. It will be assumed that this field strength also exists just above the gap, where the tape is moved across the head at a constant speed v. Then, provided that the transmit time b'/v of the tape over the gap is small compared with $1/f$, eqn. (1) represents the strength of a recording field which remains substantially unidirectional and single-valued as a given element traverses the gap. Assuming, further, that the characteristic relating the remanent intensity of magnetization and the field applied to the coating is linearized by some means, it is possible to write

$$\hat{M}_x = \eta\hat{H}_x \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

for the intensity of magnetization in an element of tape after it has left the gap, where η may be termed the 'tape sensitivity'. From eqns. (1) and (2),

$$\hat{M}_x = (4\pi N'\eta/b')\hat{I} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

It is now established practice to define the strength of a recorded signal in terms of 'surface induction', B_y, the mean magnetic induction normal to the surface of the tape in free space. If, as will be assumed in this Section, the tape thickness c is very small and the tape width w is very large compared with the recorded wavelength λ, $\hat{B}_y$ is given closely by[2]

$$\hat{B}_y = (4\pi^2 c/\lambda)\hat{M}_x \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

provided that the permeability of the tape is not much greater

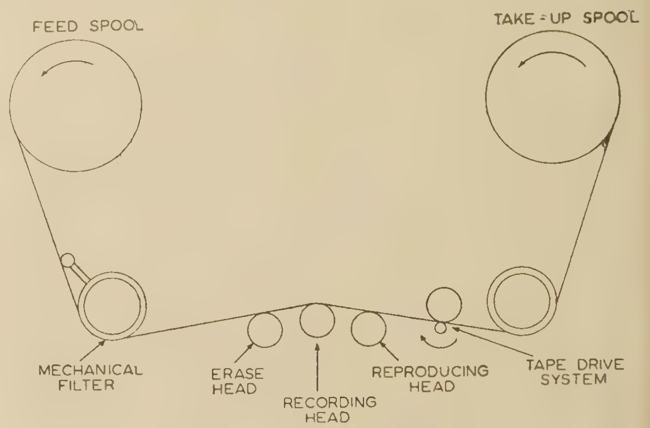


Fig. 1A.—Conventional layout of magnetic-tape recorder.

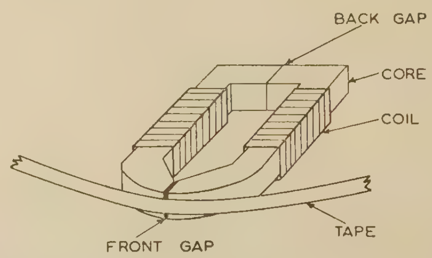


Fig. 1B.—Structure of conventional magnetic head.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Daniel was formerly with the British Broadcasting Corporation, and is now with the National Bureau of Standards, Washington, D.C., United States.
Dr. Axon and Mr. Frost are with the British Broadcasting Corporation.

than unity. From eqns. (3) and (4), the 'recording response' of the ideal system is given by

$$\alpha_i = \hat{B}_y/\hat{I} = 16\pi^3 N'\eta c/b'\lambda \quad . \quad . \quad . \quad . \quad (5)$$

Thus the recording response of the ideal system is inversely proportional to wavelength, or rises at 6 dB per octave with signal frequency.

To reproduce the signal, the tape is passed at the same speed, v, over a head similar to that used for recording, and a sinusoidally varying flux of the form $\Phi = \hat{\Phi} \sin 2\pi ft$ is established in the core which will be assumed to have negligible reluctance compared with that of the gap b. Peak flux will occur when a half-wavelength of tape bridges the gap in such a way that the flux entering each pole piece is additive. Provided that the length of the pole pieces is sufficiently great and $b \ll \lambda$, the peak value of flux is, in fact, equal to the flux emanating from the half-wavelength of tape, so that, integrating B_y between appropriate limits,

$$\hat{\Phi} = w\lambda\hat{B}_y/\pi \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The varying flux induces in a coil of N turns wound on the head an e.m.f. V of peak value given by

$$\hat{V} = 2\pi fN\hat{\Phi} = (2\pi Nv/\lambda)\hat{\Phi} \quad . \quad . \quad . \quad . \quad (7)$$

From eqns. (6) and (7), the 'reproducing response' of the ideal system is given by

$$\beta_i = \hat{V}/\hat{B}_y = 2Nvw \quad . \quad . \quad . \quad . \quad . \quad (8)$$

which is therefore independent of wavelength or signal frequency.

The 'overall response' is equal to the product of the recording and reproducing responses, and therefore should be proportional to frequency in the ideal case.

(2.2) Use of H.F. Bias

The linearity assumed in the relation between the magnetization of the recording medium and magnetizing field strength is only approximately true in practice over a limited range of unidirectional field strengths. Below this range the remanent intensity of magnetization tends to be proportional to the square of the applied field strength, and above this range it changes little with field strength as saturation is approached.

The approximately linear portion can be utilized in recording a limited range of alternating signal amplitudes if an appropriate value of 'd.c. bias' is fed to the recording head together with the signal current. A method which gives lower basic noise, and a better linearity of response over a greater range of signal amplitudes, is to replace the direct by an alternating current of similar peak magnitude and of frequency in excess of (preferably many times greater than) the highest signal frequency. No detailed explanation of the process of h.f. biasing will be attempted, but a comment on the implications of certain of its properties will be of value in discussing various phenomena in later Sections of the paper. Westmijze[3] has pointed out that the use of h.f. bias in the recording process is analogous to the method of anhysteretic magnetization discussed by Steinhaus and Gumlich.[4] In this method a linear and anhysteretic relation between remanent intensity of magnetization and unidirectional field strength applied to a specimen is obtained by superimposing on each value of unidirectional field an alternating field of high amplitude and then gradually reducing this amplitude to zero. The maximum amplitude of the alternating field is found to be unimportant, provided that it is greater than a certain value, and this is explained by making the following assumption: the final magnetic state of the specimen depends solely upon the instantaneous value of the unidirectional field when the alternating field has been reduced to a certain critical value.

The similarity of this method to the use of h.f. bias in the recording head is obvious if, for example, the longitudinal field distribution of the recording head is considered. This is such that the field strength is a maximum at the centre of the gap and decreases smoothly to zero on either side of it. Thus each element of moving tape is subjected to a maximum h.f.-bias field strength at the centre of the recording-head gap and to a gradually decreasing bias field as it leaves the gap.

In magnetic recording, however, it is observed that the recorded level falls, instead of remaining constant, as the h.f. bias current in the head is increased beyond a certain value. However, in the conventional recording head the instantaneous strength of the signal field, as well as that of the bias field, falls to zero as an element of tape leaves the precincts of the recording-head gap. Thus, when the bias current is excessive, the critical h.f. bias field strength, H_c, may be situated well beyond the trailing edge of the recording gap, where the instantaneous signal field is below the value within the gap, and the recorded level will correspond to this lower signal field.

An assumption made in the discussion of the ideal system is that the recording field is single-valued. The concept of a critical bias implies that this assumption is also valid in the practical case, since, using h.f. bias, the only effective signal field is that existing at the point where the critical bias field is located. If no h.f. bias is used, or the bias current fed to the recording head is too small, the recording field is no longer single-valued, particularly when the gap length is large compared with the wavelength. Under these conditions the magnetization of an element of tape depends not only upon the value of signal field at the critical point near the trailing edge of the gap, but upon the whole of its 'magnetic history' in traversing the gap.[5] As discussed later this can lead to effects of great complication requiring a quite separate explanation.

(3) RESPONSE AS A FUNCTION OF WAVELENGTH

(3.1) Factors Governing the Reproducing Response

(3.1.1) Gap Length.

In deducing the reproducing response [see eqn. (8)] of the ideal system it was assumed that the gap length of the reproducing head was infinitesimal. If a finite gap length is considered it has been shown[3, 6, 7] that the response is approximately given by

$$\hat{V}/\hat{B}_y = 2Nvw(\lambda/\pi b) \sin (\pi b/\lambda) \quad . \quad . \quad . \quad (9)$$

which reduces to eqn. (8) provided that $b \ll \lambda$. Within the range $b < \lambda$ eqn. (9) may be replaced by a more accurate expression, established both theoretically[3, 8] and empirically,[7]

$$\hat{V}/\hat{B}_y = 2Nvw(\lambda/\pi b) \sin (\pi b_e/\lambda) \quad . \quad . \quad . \quad (10)$$

where b_e, the 'effective gap length', is equal to the wavelength at which the first extinction is found to occur and is approximately given by $b_e = 1 \cdot 15b$.

Both eqns. (9) and (10) involve an assumption that all the reluctance of the head is contained in the gap. In practice, however, the core may have appreciable reluctance. Moreover, in order to obtain a narrow, well-defined gap at the front, a head core is usually manufactured in two halves, which are clamped or stuck together, so that a significant gap [Fig. 1B] may exist in the rear of the head. Under these conditions the reproducing response is obtained by multiplying eqn. (10) by the ratio of the front-gap reluctance S_b to the total reluctance S of the head, where

$$\frac{S_b}{S} = \frac{b/A_b}{b/A_b + l/\mu A + a/A} \quad . \quad . \quad . \quad . \quad (11)$$

and a is the length of the rear gap, A_b is the area of the front gap, A is the area of the core and rear gap and l is the length of the core of permeability μ. This factor determines the ratio of flux taking the useful path linking the coil to that taking the unwanted path across the front gap.

It is evident that the length of the front gap of a reproducing head must be chosen with two conflicting requirements in mind: the gap must be small enough to resolve the shortest wavelength, but large enough for adequate sensitivity to be achieved using practicable core materials and core dimensions. The rear gap should be made as small as possible (zero if interleaving of laminations is possible) except in the infrequent cases where reducing h.f. core losses (see Section 5) may be more important than maintaining high sensitivity. Table 1 gives the values of gap length b required to produce 6 dB loss at various frequencies and tape speeds, together with the corresponding sensitivity factor S_b/S. The figures illustrate the precision required in the manufacture of heads for short-wavelength work. In case (c) nearly all the reluctance is in the core, and of every hundred lines of flux entering the head, only about seven have a useful effect in the coils. It would be desirable to reduce considerably the reluctance around this core by reducing its length and increasing its permeability. However, in practice, a reduction of core reluctance is not easy, for in h.f. recording, ferrites with low h.f. losses, and hence low permeability, may have to be used. A more promising approach is to increase the taper at the pole tips of the head. For instance, if A_b/A is reduced to 1/10, the sensitivity in (c) is increased to 0·22.

Table 1

Example	Description	Tape speed	f	A_b/A	l	μ	a	b	S_b/S
		in/sec			in		mil	mil	
(a)	Audio frequency (laminated Mumetal)	15	15 kc/s	2/5	2	20 000	0 (interleaving)	0·53	0·93
(b)	High frequency (ferrite) ..	100	250 kc/s	2/5	1	800	0·01 (butt joint)	0·21	0·29
(c)	Video frequency (ferrite) ..	200	3 Mc/s	2/5	1	800	0·01 (butt joint)	0·035	0·065

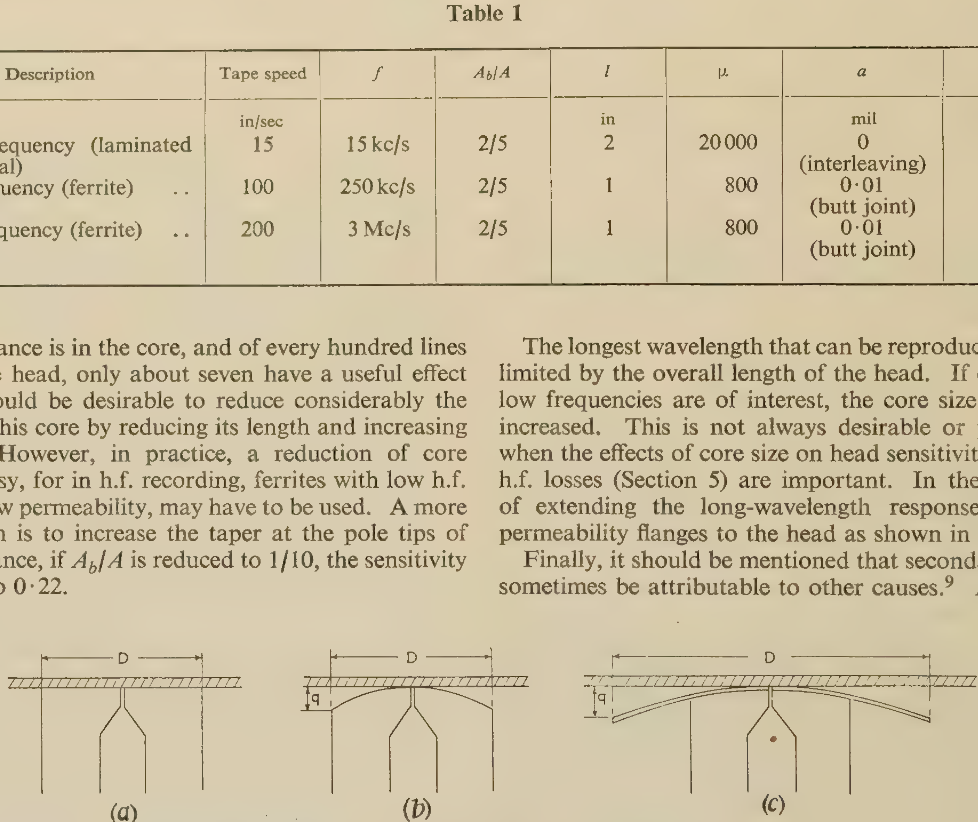


Fig. 2.—Plan views of three possible head configurations.

(3.1.2) Overall Dimensions of the Head.

Expressions (9) and (10) for the reproducing response apply only as long as the pole pieces of the head are capable of collecting all the flux available from a half-wavelength of tape in the peak-output condition. When D, the overall dimension of the head in the direction of tape travel, is less than $\lambda/2$, only a fraction $2D/\lambda$ of the available flux is collected. The extreme l.f. response thus tends to fall with decreasing frequency at a rate of 6 dB per octave. When the extremities of the head are too sharply defined in the neighbourhood of the tape [Fig. 2(a)], interference effects, commonly called 'secondary-gap' effects, arise. Provided that λ is not greater than about $3D$ the secondary-gap effect is approximately expressed by multiplying the response by the factor[3]

$$1 - 0{\cdot}205 \frac{\cos[\pi(D/\lambda + 1/6)]}{(D/\lambda)^{2/3}} \quad . \quad . \quad . \quad (12)$$

The calculated l.f. response of a head of this type is indicated by curve (i) of Fig. 3.

The undulations in the response are normally undesirable and are reduced by the more usual configuration of head shown in Fig. 2(b), in which the edges of the head are separated from the tape by a distance q. Very approximately the l.f. response of this head is obtained by modifying the factor (12) by a separation factor (Section 3.1.4) to give

$$1 - 0{\cdot}205 \exp(-2\pi q/\lambda) \frac{\cos[\pi(D/\lambda + 1/6)]}{(D/\lambda)^{2/3}} \quad . \quad (13)$$

When $q \geqslant D/2$ the response approaches the smooth curve (ii) of Fig. 3.

The longest wavelength that can be reproduced is still, however, limited by the overall length of the head. If only comparatively low frequencies are of interest, the core size can sometimes be increased. This is not always desirable or possible, especially when the effects of core size on head sensitivity (Section 3.1.1) or h.f. losses (Section 5) are important. In these cases a method of extending the long-wavelength response is to add high-permeability flanges to the head as shown in Fig. 2(c).

Finally, it should be mentioned that secondary-gap effects may sometimes be attributable to other causes.[9] At very long wave-

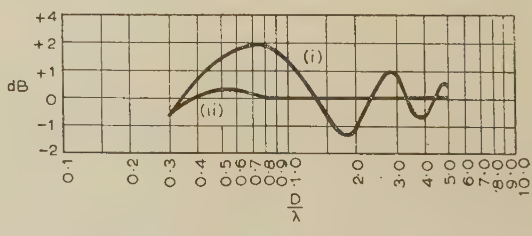


Fig. 3.—Calculated low-frequency response curves of reproducing heads.

(i) Response of reproducing head of the type shown in Fig. 2(a).
(ii) Response of reproducing head of the type shown in Fig. 2(b).

lengths, tape flux need not necessarily pass through the front surface to link with the head coils but may enter the coils by other paths. In some cases careful attention must be given to points such as the disposition of the coils and the design of screening boxes if such effects are to be minimized.

(3.1.3) **Gap Misalignments.**

The reproducing head will not work at maximum efficiency at any wavelength unless both edges of its gap are correctly aligned with the trailing edge of the recording gap. The word 'alignment' is not the happiest description of this requirement for three edges to be parallel, but it has now become familiar in this context. If the edges of the reproducing gap are parallel but at a small angle θ to the trailing edge of the recording gap, the reproducing response β_θ is approximately given by[7, 10]

$$\beta_\theta/\beta_i = (\lambda/\pi w\theta)\sin(\pi w\theta/\lambda) \quad . \quad . \quad . \quad . \quad (14)$$

where β_i is the response obtained with correct alignment ($\theta = 0$). The effect of misalignment is thus analogous to gap loss and can give rise to cyclic amplitude and phase variations in the response corresponding to an effective gap length $w\theta$.

At very short wavelengths the alignment requirements become critical. For instance, if the alignment loss is not to exceed 3 dB, the requirements for $\frac{1}{4}$ in tape are as follows:

(*a*) 15 kc/s, 15 in/sec: $\theta < 0{\cdot}1°$
(*b*) 250 kc/s, 100 in/sec: $\theta < 0{\cdot}04°$
(*c*) 3 Mc/s, 200 in/sec: $\theta < 0{\cdot}007°$

The accuracy of alignment which can be achieved, even with accurately fixed or adjustable head mountings, is limited by the degree to which the edges of the gap may be neither straight nor parallel. These errors have been discussed in detail elsewhere.[7] In addition, however, misalignment may occur in guiding the tape across the heads.[1] The tolerance in the width of $\frac{1}{4}$ in tape is normally +0, −6 mils, so that if two guides on either side of a head are situated, say, 2 in apart, a misalignment of 0·34° can exist without allowing any positive clearance in the guides at all. This would give rise to variations in head output from a maximum (at correct alignment) to zero in all the examples given above. Choice of guide spacing with reference to the variations in the tape width is obviously important.

(3.1.4) **Separation between Head and Tape.**

So far it has been assumed that the surfaces of the head and tape are in perfect contact at the front gap. However, if they are separated by a distance d, the reproducing response becomes[11] β_d, where

$$\beta_d/\beta_i = \exp(-2\pi d/\lambda) \quad . \quad . \quad . \quad . \quad (15)$$

The loss in decibels is therefore proportional to separation, and for a separation equal to a wavelength it is equal to 54·5 dB. The following examples serve to illustrate the importance of the effect:

(*a*) 15 kc/s, 15 in/sec: 5·5 dB per 0·1 mil.
(*b*) 250 kc/s, 100 in/sec: 13·6 dB per 0·1 mil.
(*c*) 3 Mc/s, 200 in/sec: 81·5 dB per 0·1 mil.

These figures indicate the great care that must be taken in the finishing of heads for short-wavelength work, so that the effective separation is reduced to a minimum. Also, of course, adequate pressure between head and tape must be provided either by a suitable combination of tape tension and lap or (less desirably) by means of pressure pads.

In some systems, such as high-speed storage drums, the heads and recording medium work out of contact to avoid wear. It is apparent that the short-wavelength performance of such devices is bound to be greatly limited by separation even of a value that would be thought extremely small in general engineering terms.

(3.2) **Factors Governing the Recording Response**

(3.2.1) **Tape Thickness.**

In many applications the shorter wavelengths encountered may be comparable with, and often less than, the thickness of the magnetic coating on the tape. This causes a departure from the ideal performance which, physically, is closely associated with the reproducing process and the exponential nature of the separation loss. When recorded level is defined in terms of surface induction, however, the effect of an appreciable tape thickness must be considered part of the recording process and the recording response is modified[11] from the ideal value α_i to a value α_c, where

$$\alpha_c/\alpha_i = (\lambda/2\pi c)[1 - \exp(-2\pi c/\lambda)] \quad . \quad . \quad . \quad (16)$$

assuming the magnetization at all wavelengths to be uniformly distributed throughout the cross-section of the magnetic coating.

At long wavelengths $\alpha_c/\alpha_i \simeq 1$ and the response approximates to the ideal which rises at 6 dB per octave with signal frequency. At very short wavelengths, however, $\alpha_c/\alpha_i \simeq \lambda/2\pi c$, implying a 6 dB per octave loss with increasing signal frequency and a consequent flattening of the observed h.f. response. For a uniformly magnetized tape of coating thickness 0·5 mil the calculated differences between the actual and ideal recorded levels at (*a*) 15 kc/s, 15 in/sec, (*b*) 250 kc/s, 100 in/sec and (*c*) 3 Mc/s, 200 in/sec amount to 10, 18 and 33·5 dB, respectively.

In effect, the surface induction at long wavelengths is established by substantially equal contributions from all layers of the magnetized coating, but the surface induction at short wavelengths is almost entirely due to a thin layer near the surface. When recording in a restricted range of short wavelengths, such as occur in some carrier systems, it may therefore be possible to reduce the coating thickness without affecting the recorded level. This may result in economy in the cost of tape and an increase in playing time for a given spool diameter.

(3.2.2) **Self-Demagnetization in the Tape.**

In addition to the useful external field a magnetic field must also exist within the recorded tape. In general, the internal field is in opposition to the magnetization creating it and may cause a reduction in the intensity of this magnetization and a corresponding reduction in surface induction.

In a recorded tape, the magnetization and demagnetizing field may have very different distributions through the depth of the tape, and a precise calculation of the self-demagnetization loss cannot easily be made. An indication of the effect has, however, been obtained[2] by calculating the mean value of the coefficient of self-demagnetization throughout the tape. For a tape magnetized uniformly in the longitudinal direction the loss is found to be negligibly small at long wavelengths, but to increase as the wavelength decreases to a limiting value $1/\mu_t$, where μ_t is the permeability of the coating, which seldom exceeds four in the commonly used materials.

In practice, the true self-demagnetization loss at short wavelengths will be considerably less than is indicated by this simple treatment. For instance:

(i) The magnetization may not all be in the longitudinal direction but may have an appreciable component perpendicular to the tape surface, the self-demagnetization of which is confined largely to long wavelengths.[3]

(ii) The longitudinal demagnetizing field is not constant throughout the depth but is much lower near the surface. This is important at short wavelengths where the useful magnetization may be confined to a surface layer.

(iii) A high-permeability reproducing-head core should reduce

the coefficient of longitudinal self-demagnetization very nearly to zero.[3] Thus, on reproduction, the estimated loss should be considerably reduced.

The last consideration provides experimental evidence of the magnitude of the self-demagnetization loss. If this were serious the recording response of a system measured by means of a conventional reproducing head and a non-magnetic conductor head[7] should be markedly different. In fact, using normal tapes, the responses are substantially the same.[12]

Many early papers attributed the major part of the h.f. loss in the magnetic recording response to self-demagnetization and specified high coercivity as the controlling magnetic property in obtaining a good h.f. response. It now appears that these conclusions were inaccurate: self-demagnetization probably contributes only a small part of the total loss, and the relevant magnetic property is the permeability, rather than the coercivity, of the tape material.

(3.2.3) Non-Uniform Recording Field.

So far, it has been assumed that the recording field is uniformly distributed throughout the depth of the tape. In most cases, however, the recording field strength H will decrease appreciably with distance y from the surface of the head. If H_0 is the strength of the field within the gap, the maximum strength H_x of the longitudinal field outside the gap is approximately given by[2, 13]

$$H_x/H_0 = (2/\pi) \arctan (b'/2y) \quad . \quad . \quad . \quad (17)$$

It follows from eqn. (17) that a gap length b', small compared with the tape thickness c, will lead to a marked decrease in the signal and bias field strengths through the tape, but for other reasons, to be given in the following Sections, it is undesirable to make the gap length very large. Often b' is made approximately equal to the tape thickness, normally 0·5 mil, but the longitudinal field strength at the base of the coating will, even so, be less than one-third of that at the surface.

The decrease of the signal field strength through the depth of the tape is less important than the decrease of the bias field strength, for the latter makes it impossible to bias correctly the whole of the coating. With serious non-uniformity of field, two extreme cases arise as follows:

(*a*) The bias may be adjusted so that the outer layers are correctly biased. In this case the inner layers may be grossly under-biased so that the recording on them is very non-linear, and, since the signal field strength has also fallen, of low level. This will not greatly affect short wavelength work, where the useful magnetization is confined to the surface layers, but it may greatly detract from the performance at long wavelengths ($\lambda > c$).

(*b*) The bias may be adjusted so that the inner layers are correctly biased. In this case the outer layers may be grossly over-biased, leading not so much to distortion in these layers as to a reduction in their intensity of magnetization. The effect is contrary to the previous case in that a greater loss will now occur at short wavelengths for which the outer layers are relied upon for the major contribution to surface induction. It will explain, in part, the fact that short wavelengths are more easily over-biased than long wavelengths, particularly when the tape thickness is large compared with the gap length of the recording head.[2] Over-biasing is further discussed in Section 3.2.4.

To summarize, the recording gap length should be as large as is permitted by the considerations to be discussed in Sections 3.2.4 and 3.2.5. Only if attention is strictly confined to very short wavelengths will a short gap length be entirely advantageous.

(3.2.4) Rate of Extinction of Recording Field.

In discussing h.f. bias in Section 2.3 it was suggested that, in conditions of adequate biasing, the final remanent magnetization of an element of tape leaving the recording head is established at the point where the bias field strength has fallen to a critical value H_c. This might indeed be precisely the case for an element of tape consisting of a single particle or magnetic domain. In general, however, all the particles in the coating will not be identical in form and cannot be expected to require identical values* of H_c. Indeed, as pointed out by Westmijze,[3] if there were a unique value of critical field the sensitivity/bias curve should rise vertically to its maximum value instead of rising, in practice, with a finite (although steep) slope.

This modification of the hypothesis will not materially affect the linearizing action of the bias. It will, however, affect its second attribute, namely the creation of conditions in which the recording signal field experienced by a given element of tape can be regarded as single-valued. Thus if the values of H_c required by the domains in an element of tape are distributed between H_2 and H_1, recording in the element must take place over a distance ξ in which the bias field strength falls from H_2 to H_1. Ignoring the fall in instantaneous signal field strength over this distance, the effect is obviously analogous to an aperture loss of the form

$$\alpha_\xi/\alpha_i = (\lambda/\pi\xi) \sin (\pi\xi/\lambda) \quad . \quad . \quad . \quad . \quad (18)$$

which will have no appreciable effect on the recording response at long wavelengths, but will cause a serious loss when λ becomes comparable with ξ. In practice, of course, the signal field strength will fall to the same extent as the bias field strength over the critical range, and the distribution of H_c may be complex. Nevertheless, it is of interest that series of minima can be observed in the recording response at short wavelengths when high bias levels are used. These minima appear to be unrelated to known gap phenomena, and their position is found to depend very much upon bias level.

The distance ξ will depend upon the rate of extinction of the recording field, and if this could be made instantaneous ξ would be zero. On the other hand, if the decay of field were very slow compared with the period of the signal, the combined signal and bias recording fields might act simply as an erasing field and the tape would emerge from the recording head in a neutral condition. In practice, therefore, the rate of extinction should be made as high as possible—a conclusion supported by the experiments of Muckenhirn.[15] Even with very short gap lengths, however, the rate may still be significant, and a substantial part of the h.f. loss of biased systems can probably be attributed to this 'critical range' effect. Alternatively, of course, the distance ξ is reduced when the range $(H_2 - H_1)$ of critical fields required by the coating particles becomes smaller. If all particles required the same critical field, H_c, the rate of extinction of the recording field would be unimportant and no h.f. loss should result from it.

A secondary effect of considerable importance is that the distance ξ will also depend upon bias current. This arises because the decay of the bias field on leaving the gap is not uniform. If the bias current in the head is increased, the limits of the critical range are moved further away from the gap and are spaced further apart on a more gradually sloping part of the bias-field decay curve. This may account for the fact that short wavelengths are found to be more easily over-biased than long wavelengths, even when recording gaps substantially greater than the tape thickness are used and the effects noted in Section 3.2.3 cannot play a large part. In practice, the short-wavelength over-biasing phenomenon often imposes a serious limitation on the performance of biased systems owing to the compromise adjustment which has to be made to avoid serious h.f. loss on the one hand and l.f. insensitivity and distortion on the other.

(3.2.5) Interference Effects in the Recording Gap.

If good linearity of response is required the use of h.f. bias in recording is essential. The discussion in the preceding two Sec-

* The theory of magnetization in heterogeneous alloys developed by Stoner and Wohlforth has been shown by Osmond[14] to be applicable to tape materials. The existence of critical fields is then readily explained in terms of particle-shape anisotropy.

tions indicates, however, that there are also certain disadvantages, particularly when short wavelengths are to be recorded. In applications where a good response at very short wavelengths is a more important requirement than linearity, such as computer stores, an unbiased system is frequently employed.

It has been shown,[5] however, that the recording field affecting the tape when h.f. bias is absent is not single-valued, and serious interference effects may then occur in the recording gap. These give rise to undulations in the recording-frequency characteristic which are related to the finite length of the recording gap, with minima occurring when the gap length b' is approximately equal to $3\lambda/4$, $7\lambda/4$, etc. In pulse recording work the phase-distortion aspect of these undulations may be of greater importance than the amplitude variations. It has also been shown[5] that the effect of adding and increasing bias is gradually to eliminate the minima, but, unfortunately, they cannot always be entirely removed before an appreciable over-biasing of the shortest wavelength occurs owing to the secondary effects discussed in Sections 3.2.3 and 3.2.4. The interference phenomena in the recording process may therefore also be significant in systems in which, perhaps to avoid h.f. over-biasing, too low an h.f. bias is used in conjunction with an appreciable gap length.

A conclusion of considerable importance to draw from this discussion is that when, for various reasons, a recording system is to be used with little or no h.f. bias, the gap length of the recording head should be small enough for the condition, say, $b' < \lambda_s/2$ to obtain, where λ_s is the shortest wavelength to be recorded.

(3.2.6) **Separation between Head and Tape.**

The surface of the tape may be intentionally or accidentally separated from the surface of the recording head for reasons similar to those described in the reproducing head case but with results which although more complex, are generally less serious. Two principal effects can be envisaged as follows:

(i) Separation will decrease the strength of both the signal and bias fields and cause a general loss in recording response. The effect will be most marked if, when in contact, the bias field strength is only just sufficient to give maximum sensitivity.

(ii) Separation will cause a lower rate of extinction of the recording field, since the leakage flux is more widely spread over planes removed some distance from the gap. This, from the arguments of Section 3.2.4, may cause a loss at shorter wavelengths, but the loss is normally very small compared with that arising in an equivalent separation from the reproducing head.

(4) AMPLITUDE AND SPEED FLUCTUATIONS

(4.1) Effect of Amplitude Fluctuations

In practice, the e.m.f. in the reproducing head contains components of noise, forming part of the phenomenon known as 'modulation noise', owing to undesired amplitude fluctuations which are superimposed on the input signal in the recording and reproducing processes. A typical distribution of the instantaneous amplitudes in the (nominally constant) envelope of a 10 kc/s tone, recorded and reproduced at 15 in/sec on a production tape of average quality, is shown in Fig. 4. Analysis of these fluctuations indicates that they are largely random so that they contribute a true noise voltage to the input signal and so reduce the signal/noise ratio. In the following Sections some consideration will be given to the origins of these fluctuations in order to assess how far they may be reduced by careful design and manufacture of the components of the system.

(4.2) Factors Causing Amplitude Fluctuations

(4.2.1) **Particle Nature of the Coating.**

In most modern systems the recording medium consists of small particles of a suitable magnetic oxide of iron which are

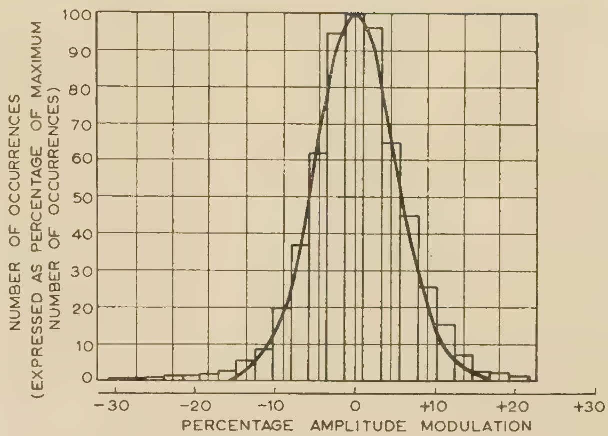


Fig. 4.—Distribution of amplitudes in modulation envelope of recorded 10 kc/s tone.

Period of observation: 10 sec.

mixed with a suitable binding agent and coated on to a plastic tape or other surface, as required. The medium may therefore consist of crystals which vary, according to the manufacturing process, in size, shape and orientation[14] and in separation from one another.[16] The uniformity of distribution and size of the particles, and of their orientation, is, of course, vastly increased in modern tapes compared with the products of a few years ago. Nevertheless it remains true that the variation of magnetization within any one recorded wavelength must occur in a series of discrete steps, the magnitude of which depends on uniformity in the factors noted above. The size of single domains imposes an ultimate limit on short-wavelength recording, for, clearly, no frequency corresponding to a half-wavelength shorter than the average size of domain can have a sensible effect on the medium. This ultimate limit cannot be achieved, however, whilst the particles, even if single domains, are separated by some finite distance owing to the presence of the binding agent. The practical limit is then larger by a factor depending on the average separation of particles, i.e. the 'packing factor'. A random variation in packing factor will create an even greater limitation, since the random variation of amplitude which it imposes on the recorded signal appears as noise in the reproduced output. These coating variations will particularly affect the signal/noise ratio in the short-wavelength region, in which only a surface layer of the tape is employed. The effect at longer wavelengths should be less severe if the lack of uniformity in size, packing and orientation is the same at all depths within the coating, for the greater the number of layers contributing to the output the less will be the relative effect of a variation of signal in any one of them.

(4.2.2) **Irregularities in the Surface of Backing.**

Another noise contribution of this type occurs when the surface of the backing on which the medium is supported is irregular. Lack of smoothness in the backing creates a gross variation in the dispersion of the particles at the base of the coating, and this will assume importance when the under layers are significant in the recording and reproducing processes. This is the case at long wavelengths if the signal and bias fields from the recording head are such as to establish an appreciable magnetization at the base of the coating. This effect accounts for the comparatively high modulation noise which occurs in tapes with paper backing, compared to those with plastic backing, the fibrous nature of paper making it difficult to attain the same degree of surface smoothness that can exist on a plastic material.

(4.2.3) **Variations of Contact between Head and Tape.**

The level of signal appearing at the terminals of the reproducing head has been shown to depend on the effective separation of the recording medium from the heads in the recording and reproducing process. In practice, undesired changes in the separation of the tape from the heads will occur for several reasons and create another source of modulation noise.

The lack of uniformity in the coating, previously discussed, also implies a lack of perfect flatness at the surface, so that the effective separation of the tape from the heads must vary in a manner depending on the surface variations. Extreme cases of the latter can lead to an almost complete disappearance (a 'drop-out') of the signal. These variations may be reduced by suitable polishing of the tape surface after manufacture. Similar fluctuations may also arise when tape debris, which may be created by rough head or guide surfaces, is carried over the heads. In this connection, a build-up of static charge on the tape surface, which retains debris or attracts dust, should be avoided.

Tape-width variations may also cause fluctuations when, in order to obtain high accuracy of alignment, the clearance in the guides is reduced too much. If a portion of tape occurs which is outside the tolerance allowed, it will bend, and be lifted from the head surface, in passing through the guides. Inaccuracies in the driving system, which vary the tape tension, will also vary the effective separation of head and tape.

(4.3) Effect of Speed Fluctuations

In practice, even in a well-designed system, the tape speed is never perfectly correct for an appreciable period, and the reproduced signal contains the effects, in the form of frequency modulations, of errors in both the recording and reproducing processes. The errors are attributable to mechanical imperfections of the drive system and to interactions between the moving tape and the transport system. Certain well-defined frequencies of modulation may be present, but it has been shown elsewhere[17] that a large random element may also be observed. As such, the speed changes make an addition of modulation noise to the signal and reduce the ability of the system to provide accurately timed information when required. In the next few Sections the origin of these speed changes will be examined and measures which may be taken to reduce them will be briefly discussed.

(4.4) Factors Causing Speed Fluctuations

(4.4.1) **Tape Transport System.**

In the tape-recording system illustrated in Fig. 1A the drive is obtained from a rotating capstan against which the tape is held by a spring-loaded rubber idler. This system is widely employed, and it is unnecessary to describe here the manufacturing accuracy and measures required if adequate freedom from speed fluctuations is to be obtained. In many cases, however, insufficient attention is paid to the spooling system, although it is obvious that any sudden fluctuations in spooling torques must vary the tension and the velocity of the tape over the heads. Fluctuations of this kind arise if there are eccentricities in the spools, if the tape comes into accidental contact with the sides of the spool or if there are inaccuracies in any pulleys which guide the tape to the heads. The tape over the heads is, of course, in some degree isolated by the capstan system from sudden torque changes in the take-up-spool system. The same is not true of the feed-spool system, however, and some other measure, such as the mechanical filter indicated in Fig. 1A, may be necessary. Where high mean accuracy is required it is possible to control the torque of the spooling motors by a servo system actuated by tension indicators which are placed in contact with the tape adjacent to the spools.[18] Comparative isolation from both take-up and feed spools is also possible in an alternative drive system[19] illustrated in Fig. 5. The tape is driven by the capstan at two points, A and B, and passes over a pulley C between them, the recording and reproducing heads being mounted at convenient points inside the loop so formed.

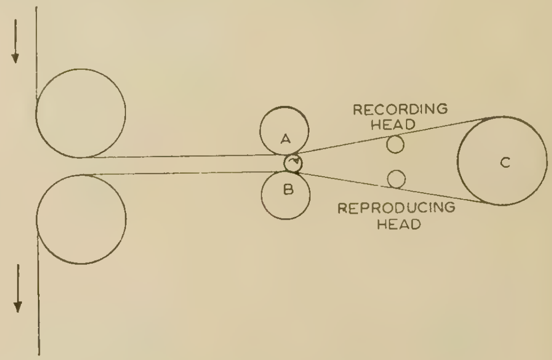


Fig. 5.—'Double-pinch' driving system.

(4.4.2) **Interaction between Tape and Transport System.**

The small variations of width which may exist along the length of the tape are also important as a source of speed fluctuations. Completely accurate guiding requires the guide width to be identical to the nominal width of the tape, but, in practice, some tolerance must be allowed. When a length of tape occurs which is wider than the guide the frictional forces between the edges of the tape and guide become large. On the other hand, when a length of tape occurs which is narrower than the guide the edge frictional forces are confined to one edge only or may even be absent altogether. If there is random variation in the width of the tape an appreciable random component must be present in the frictional forces. When the length of tape over the heads is not isolated from these forces its velocity must vary accordingly.

In the conventional form of static guide the surface of either the backing or the coating also comes into contact with the face of the guide and another set of frictional forces occurs. A similar set is also generated as the tape passes each head. Random surface conditions on either coating or backing (Section 4.2.3) will vary these forces to introduce another random element into the system. Surface friction in the guides will be eliminated if they are so designed and positioned that the tape comes into contact with them only at its edges or if the guides take the form of pulleys which are driven by the tape.

Surface friction is also significant in the origination of another distinctive set of higher-frequency speed fluctuations, which may be observed under critical conditions. These fluctuations, which have been investigated by Werner,[20] result from a longitudinal oscillation of the tape generated by static elements such as the heads, the excitation being provided by the frictional forces. If no slip is occurring at, for example, the guide pulleys situated before the heads and the driving point on the capstan which follows them, these elements have zero speed relative to the tape and they constitute bridges or nodes between which the oscillations are propagated. If s is the distance between the nodes, the frequency of the oscillation is given by $f = (1/2s)\sqrt{(E/\rho)}$, where E is the modulus of elasticity and ρ is the density of the tape. In practice, the frequency of the oscillations is found to lie between 1 and 3 kc/s, and their amplitude varies with the coefficient of friction of the tape and the tape tension, the latter determining the pressure on the heads and hence the frictional forces set up. The effect is therefore particularly large in machines in which contact between the heads and tape is maintained by pressure pads. The

mplitude will be reduced in a conventional system if the follow-ng requirements are observed:

(a) The tape tension is reduced as far as possible, consistent with adequate contact.
(b) The coefficient of friction of the tape is made as small as possible.
(c) Fixed guides touch the tape only at its edges.
(d) The head surfaces are extremely well polished.

These factors cannot usually be adjusted to eliminate the oscilla-ions completely, but the traces which remain may be further reduced by 'loading' the tape, at convenient points between the heads, with further pulleys on which no slip occurs.[20] It is also beneficial to mount the heads near the drive system and to make the distance between heads as small as possible. These latter measures tend to generate the oscillations at higher frequencies at which they are more severely attenuated. When loading pulleys are introduced they must be of high accuracy, in order to ensure that they do not introduce periodic fluctuations on their own account.

(5) LOSSES IN THE HEAD CORE

(5.1) Nature of Core Losses

The sensitivity of a reproducing head is proportional (Sec-tion 3.1.1) to the ratio of front-gap reluctance, S_b, to total reluctance, S. The sensitivity of a recording or an erasing head may be defined in terms of the field strength produced in the gap for a given current in the coils, and is then proportional to $1/S$. Thus, for all types of heads, the ratio of the sensitivity σ_f at a frequency f to the sensitivity σ_0 at zero or very low frequencies is given by

$$\frac{\sigma_f}{\sigma_0} = \frac{S_0}{S_f} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

This assumes that the reluctance of the gap remains constant but that the total reluctance of the head increases from S_0 to S_f as the frequency is raised from zero to the value f. Such an increase is attributable to the fact that an increasing amount of energy is absorbed in eddy-current, hysteresis and 'residual' losses in the core as the frequency is raised. These losses are accompanied by a lag in phase and can be described by con-sidering the effective permeability of the core to change from a real value μ_0 at zero frequency to a complex quantity μ_f at a frequency f, where

$$\mu_f = \mu' - j\mu'' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

If the head has a front-gap length b, a rear-gap length a, a core length l, and a cross-sectional area A throughout, eqn. (19) can be written

$$\frac{\sigma_f}{\sigma_0} = \frac{b + a + l/\mu_0}{b + a + l/\mu_f} = \frac{\gamma + 1/\mu_0}{\gamma + 1/\mu_f} \quad . \quad . \quad . \quad (21)$$

where
$$\gamma = \frac{b + a}{l} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

The core is often tapered towards the front gap when the cross-sectional area A_b of the front gap is considerably less than A. This can be taken into account by replacing b by bA/A_b in eqn. (22).

Substituting for μ_f in eqn. (21) and rationalizing,

$$\left|\frac{\sigma_f}{\sigma_0}\right| = \frac{\gamma\mu_0 + 1}{\mu_0}\left[\frac{(\mu')^2 + (\mu'')^2}{(\gamma\mu' + 1)^2 + (\gamma\mu'')^2}\right]^{1/2} \quad . \quad . \quad (23)$$

and the angle Φ of the phase lag is given by

$$\tan \Phi = \mu''/[\gamma(\mu'^2 + \mu''^2) + \mu'] \quad . \quad . \quad . \quad (24)$$

From eqn. (23) the head sensitivity remains substantially constant provided that $\gamma\mu' \gg 1$, i.e. the reluctance of the core is always small compared with that of the gaps. In general, therefore, the effect of core losses can be minimized by (a) making the front gap length as large as possible without causing severe interference effects at the shortest wavelength, (b) introducing an appreciable gap in the rear of the head, if absolute sensitivity requirements allow it, and (c) making the core as short as possible and of a material of high permeability and low loss over the required frequency range. The effect of core losses is reduced by (b) at the expense of a reduction in absolute sensitivity equivalent to multiplying the response when $a = 0$ by the factor

$$\frac{b + l/\mu_0}{a + b + l/\mu_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

Loss of sensitivity in recording heads can be offset by increased recording currents, and here the larger rear gap may be well worth while, for, in addition to reducing the effect of core losses, it also tends to reduce the effects of non-linearity in the core. In reproducing heads, however, absolute sensitivity is usually of overriding importance and a large rear gap is unacceptable. The overall sensitivity is then proportional to the product of eqns. (23) and (11).

(5.2) Measurement of Core Losses

Two methods of measuring the effect of core losses on head performance are in common use, particularly in connection with frequency-characteristic standardization. The first, and perhaps the more direct, method relies upon separating the losses that fundamentally depend on frequency from those that fundamentally depend on wavelength by taking measurements of overall response at a variety of tape speeds. In the second method an alternating flux is induced in the head from a small conducting loop or coil placed near the front gap, and core losses are determined by measuring the departure of the e.m.f. generated in the head coil from a 6 dB per octave law.

In many cases, however, it is convenient to assess the core losses from bridge measurements of the impedance of the head at various frequencies. When L_0 and R_0 are the series inductance and resistance at zero frequency and L_f and R_f are the effective values measured at a frequency $f = \omega/2\pi$, the effect of core losses on head sensitivity is given by

$$\left|\frac{\sigma_f}{\sigma_0}\right| = \left|\frac{S_0}{S_f}\right| = \frac{L_f}{L_0}\left(1 + \frac{1}{Q_f^2}\right)^{1/2} \quad . \quad . \quad . \quad . \quad (26)$$

where
$$Q_f = \frac{\omega L_f}{R_f - R_0} \quad \text{and} \quad \tan \Phi = \frac{1}{Q_f}$$

(5.3) Losses in Laminated Alloy Cores

The most frequently used core materials, particularly in the lower-frequency ranges, are nickel–iron alloys. These materials have extremely high permeabilities but comparatively low resistivities, so that normally they are laminated to avoid excessive eddy-current losses. For a lamination thickness δ and resistivity ρ the eddy-current loss gives rise to a complex permeability[21] in which

$$\left.\begin{aligned} \mu' &= \frac{\mu_0}{\psi}\frac{\sinh \psi + \sin \psi}{\cosh \psi + \cos \psi} \\ \mu'' &= \frac{\mu_0}{\psi}\frac{\sinh \psi - \sin \psi}{\cosh \psi + \cos \psi} \\ \psi &= 2\pi\delta\sqrt{\frac{\mu_0 f}{\rho}} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (27)$$

In most cases the eddy-current loss will be much greater than the hysteresis or residual loss, even when considering recording or erasing heads in which the maximum flux density may be high. Eqns. (27) may therefore be used in conjunction with eqn. (23) to give a close estimation of the head sensitivity loss using a laminated material. For values of ψ greater than 5, $\mu' \simeq \mu'' \simeq \mu_0/\psi$ and eqn. (23) becomes

$$\left|\frac{\sigma_f}{\sigma_0}\right| = \frac{(\gamma\mu_0 + 1)\sqrt{2}}{[(\gamma\mu_0 + \psi)^2 + (\gamma\mu_0)^2]^{1/2}} \quad . \quad . \quad . \quad (28)$$

Thus when $\psi \gg \gamma\mu_0$ the sensitivity of a laminated head will tend to fall at a rate proportional to the square root of frequency and the phase lag will tend towards 45°.

As an indication of the lamination thickness required, consider a possible head in which $b = 0 \cdot 3$ mil, $l = 1 \cdot 5$ in and the halves of the core are interleaved to avoid a rear gap. Let the laminations be such that $\rho = 40$ microhm-cm and $\mu_0 = 10\,000$, so that $\gamma\mu_0 = 2$. Then if an eddy-current loss of 6 dB is acceptable at 16 kc/s the lamination thickness must not exceed 6 mils, and for the same loss at 600 kc/s it must not exceed 1 mil.

In practice the improvement obtained with a lamination thickness less than 5 mils will seldom be as great as that expected from simple theory. Additional losses, possibly due to skin effects, become significant and the laminations are difficult to prepare. Mechanical working after the final heat treatment, a certain amount of which is usually unavoidable in head manufacture, may also cause large local reductions in the effective permeability.

(5.4) Losses in Ferrite Cores

Eddy-current losses can be reduced to negligible proportions if the head core is made of a ferrite material, the resistivity of which may be a million or more times that of the common alloys. Several grades are available covering a wide range of permeability, residual loss coefficient, maximum flux density and coercivity. In general, the permeabilities and maximum flux densities are much lower, and the coercivities are much higher, than those of magnetic alloys.

In reproducing heads the desirable properties of the core are high permeability and small residual loss. When information is available on the variation of the components of the complex permeability with frequency the most suitable material for a particular reproducing head can be chosen by calculating sensitivity/frequency curves from eqn. (23). It is usually found that ferrites of permeability much less than 1 000 give an unacceptably poor sensitivity at low frequencies, even though they may have very small losses, so that the number of grades to choose from is limited. With available ferrites of permeability of the order of 1 000 it is possible to produce a reproducing head with negligible loss up to a frequency of approximately 1 Mc/s.

The core of a recording head should also have high permeability and low residual loss, but here the hysteresis loss must be considered. Even without bias the flux density in the core may be high enough for hysteresis losses to become appreciable at the higher signal frequencies, and a ferrite of low coercivity should therefore be used. When bias is used the recording head must, in effect, be designed to handle a considerable power at a frequency normally many times greater than the highest signal frequency. Hysteresis effects at the bias frequency can cause excessive losses, which, since ferrites are usually poor thermal conductors, may give rise to over-heating. In extreme cases, when out-of-contact working or high-coercivity tape necessitate very high bias currents, the temperature of the core may rise above the Curie point, which is of the order of 150° C. It may also be found in such cases that the maximum flux density of certain grades of ferrite is inadequate. When bias is necessary and these considerations limit its frequency to a value f_b, the highest signal frequency that can satisfactorily be recorded is also limited to a value f_b/K, where K ought to be at least 3. If K is less than 3 the biasing action becomes less efficient, the signal level falls and non-linearity may be present. Unwanted frequency components may then be generated within the signal band by interaction between signal and bias.

Similar considerations affect the choice of core for the erasing head, in which the frequency must usually be well above the highest signal frequency that could be recorded by the erasing head if it were used as an unbiased recording head.

(6) SOME PROBLEMS OF HEAD CONSTRUCTION

(6.1) Construction of Ferrite Heads

The use of ferrites introduces new problems of construction and finish, some of which are very difficult to solve. Three principal difficulties arise as follows:

(*a*) Ferrites are difficult to mould in small intricate shapes owing to their severe contraction after sintering.
(*b*) They are brittle and difficult to work.
(*c*) The surface finish obtainable is limited by the existence of small holes and fissures in the material.

The first two difficulties can be largely overcome by developing suitable grinding techniques. The third is more difficult to surmount and attempts made so far have not met with a great deal of success. It is, indeed, possible to assemble a ferrite head, polish the front surface, and obtain a finish which looks smooth under the microscope. When the head is used, however, it is inevitably found that the performance rapidly deteriorates; tape dust piles up on the contact surface, and, after cleaning and re-examination, the finish is found to be rough. These findings have been confirmed by other workers.[22]

Until such time as a more suitable ferrite material is developed, a method[23] of overcoming this difficulty which can be made to give satisfactory results consists of facing a ferrite core with a thin sheet of high-permeability alloy. The technique is essentially one of combining the good surface-finish and wearing properties of the alloy with the low core losses of the ferrite. The alloy is cemented to the two halves of the core before the gap faces are polished and the head is assembled. The alloy sheet should, of course, be as thin as possible to avoid eddy-current effects and should be in intimate contact with the core proper to avoid excessive loss in general sensitivity. An advantage of the technique is that, by sharply tapering the ferrite pole-tips below the alloy face and by using an alloy of permeability much greater than that of the ferrite, an effective value of γ can be obtained which is considerably greater than b/l for the ferrite alone.

(6.2) Head Systems for Multi-Track Recording

In some applications of magnetic recording, several signals are required to be recorded simultaneously, and this has led to the development of multi-track systems in which the various input signals are handled by separate pairs of recording and reproducing heads on parallel tracks on the medium. In some arrangements the individual heads are manufactured separately and laid out in echelon formation across the width of the tape or drum. Advantages of this are that direct crosstalk between the various heads is eliminated and the alignment of each reproducing head can be precisely adjusted to match its associated recording head. Difficulties arise, however, when great phase accuracy is required between the various reproduced signals, and complicated mechanical adjustments have then to be provided to make the various head spacings equal and maintain them so under conditions of mechanical movement, tape stretch or temperature change. An

alternative method is to manufacture the heads, with screens between them, in a composite stack and ensure, by a suitable manufacturing process, that all the gaps are in perfect alignment both with respect to one another and to the direction of tape travel. This becomes more difficult as the number of heads in the stack is increased and the permissible time or phase error is decreased. A common method is to manufacture the stack of heads in two halves with respect to a line through the centre of the gaps (e.g. see Reference 24), but little information has been revealed on detailed procedures.

The magnetic screens which are placed between adjacent heads to prevent direct crosstalk do not reduce appreciably the inter-track crosstalk which arises at long wavelengths as flux from one recorded track spreads into adjacent reproducing heads. If the magnitude of this type of crosstalk (which is common to both stack and echelon systems) is unacceptable, the track separation must be increased or a carrier system adopted in which the order of recorded wavelengths is shorter and the spread of flux correspondingly less.[18]

(7) EQUALIZATION OF THE RESPONSE

In general, the overall response, V/I, of a magnetic recorder at first rises, at a rate approximately proportional to frequency, to a maximum and then falls with increasing frequency owing to the combined effect of the various losses detailed in previous sections. Let it be assumed that the inherent differentiation of the system has been corrected by inserting a suitable integrating stage into the reproducing amplifier. The amplitude/frequency characteristic is then (ignoring very-long-wavelength losses) essentially that of a low-pass network which can theoretically be corrected by inserting the appropriate high-pass network in the chain. The total amplitude correction possible, and the relative amounts which are placed in the recording and reproducing chains, depends on the frequency spectrum of the input signal and the noise and distortion characteristics of the system.

Equalization of phase as well as amplitude may be more difficult, however, depending on the nature of the losses. In so far as the amplitude and phase relationships of these losses are approximately similar to those which occur in linear, passive, electrical networks, their correction, theoretically at any rate, is straightforward. Core losses in the heads can, for example, be corrected in phase and amplitude by comparatively simple methods. Thus, if laminated recording and reproducing heads are used, in which eddy currents cause the major loss, their combined loss is eventually proportional to frequency and the eventual phase lag is 90°, which, fundamentally, can be corrected by means of a simple RC network. Another technique, applicable in simple form to the correction of recording-head core losses, is to apply a negative-feedback voltage, derived from the integrated e.m.f. from a secondary winding on the head, to the input of the recording amplifier. If sufficient gain is inserted in the feedback loop a linear relation between head flux and current can be obtained regardless of core losses. A similar, though more complicated, procedure may be used to correct losses in a reproducing head.

However, the various other losses in the system do not have a parallel in linear passive networks. Thus aperture effects are purely amplitude losses which introduce no distortion of phase, unless the response actually contains a reproducing-gap minimum or recording-gap interference effects are present. Similarly the separation and the tape-thickness losses are attenuations which increase, without any associated phase change, with decreasing wavelength. These losses may be equalized using derivative equalizer techniques[25] in which only even-power derivatives of frequency are involved. In practice, the degree to which the equalization curve required may be approximated is limited by the highest-power derivative with an acceptable signal/noise ratio which is available from the equalizer. The derivative method has the advantage that adjustments can be made while the system is actually operating and, by introducing odd-order derivatives (the first is usually available in the direct output from the reproducing head), correction can be made at the same time for the losses in the head cores and in the associated electrical equipment. It is not easy to carry out phase measurements on recording equipment, owing to the fact that there is a large (indefinitely large on separate record and replay) delay between output and input, and, inevitably, a certain amount of wow and flutter. The only possible techniques are similar to those used on long lines or radio links when the input signal is not available for comparison with the output.

(8) DISCUSSION

The fundamental inefficiency of the conventional system at long wavelengths may be overcome by the adoption of a modulated carrier system which utilizes a band of frequencies, the wavelengths of which lie outside the inefficient region. The severity of the unwanted amplitude modulation which occurs in practice will be increased at the shorter wavelengths of the carrier system, and this may prohibit the use of an amplitude-modulated carrier. This difficulty may be overcome by frequency modulation, which is also the solution when a reduction of the unwanted amplitude modulation in conventional recording is a requirement. Stringent requirements are then imposed, however, on the speed constancy of the system. In all cases the recording speed necessary will be many times that required to record the highest modulating frequency conventionally, so that all carrier systems are less economic in the use of tape. They are, however, inherently more free from the effects of accidental printing, for the wavelengths on the tape are usually much less than that at which optimum printing[26] occurs.

In considering the use of carrier modulation to overcome long-wavelength difficulties it must be remembered that it is not always necessary to adopt a carrier method, and, in some cases, where very low frequencies are involved, a flux-sensitive reproducing head[27, 28] may be used which allows the recording chain, and the tape speed, to remain unchanged.

In general, however, it appears that the most fruitful advance towards greater utility of the magnetic system is an improvement in its high-frequency and short-wavelength performance. Improvement in this respect will extend both direct and carrier recording applications. The developments required are partly of a magnetic and partly of a mechanical nature.

In heads the improvements required are mainly in the core materials. The laminated alloy cores employed for audio frequencies are unsuitable for high frequencies owing to the severity of the eddy-current losses. Although the ferrite cores available are better in this respect, they are inferior in most other respects. First, lower hysteresis and residual losses are desirable to maintain the signal strength at high frequencies and reduce the heat generated. Secondly, much greater mechanical stability and much better machining properties are necessary to allow the manufacture of fine gaps and intimate contact surfaces.

The magnetic properties of the tape coating should be such that a high maximum sensitivity is obtainable at a low bias field, and for short-wavelength applications its various components should require similar values of critical field H_c, to provide permanent magnetization. The coating should possess a high uniformity of distribution and a surface smoothness which gives low friction and good intimacy of contact with the heads. The mechanical properties of the backing are almost equally important: it should be smooth, flexible, free from appreciable stretch and stable under normal variations of temperature and

humidity. The width of the tape should be maintained within the closest tolerances so that the guiding system may be made to the high accuracy required to eliminate alignment errors.

(9) ACKNOWLEDGMENT

The paper is published by permission of the Chief Engineer of the British Broadcasting Corporation.

(10) BIBLIOGRAPHY

(1) Selsted, W. T., and Snyder, R. H.: 'Magnetic Recording—A Report on the State of the Art', *Transactions of the Institute of Radio Engineers*, Professional Group on Audio, 1954, p. 137.

(2) Daniel, E. D.: 'The Influence of some Head and Tape Constants on the Signal Recorded on a Magnetic Tape', *Proceedings I.E.E.*, Paper No. 1507 R, May, 1953 (**100**, Part III, p. 168).

(3) Westmijze, W. K.: 'Studies on Magnetic Recording', *Philips Research Reports*, 1953, **8**, pp. 148, 161, 269 and 343.

(4) Steinhaus, W., and Gumlich, E.: *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 1915, **17**, p. 369.

(5) Axon, P. E.: 'An Investigation into the Mechanism of Magnetic-Tape Recording', *Proceedings I.E.E.*, Paper No. 1149 R, May, 1952 (**99**, Part III, p. 109).

(6) Lübeck, H.: 'Magnetische, Schallaufzeichnung mit Filmen und Ring-Köpfen', *Akustische Zeitschrift*, 1937, **2**, p. 273.

(7) Daniel, E. D., and Axon, P. E.: 'The Reproduction of Signals Recorded on Magnetic Tape', *Proceedings I.E.E.*, Paper No. 1499 R, May, 1953 (**100**, Part III, p. 157).

(8) Schmidbauer, O.: 'Das Feld des Harmonisch Magnetisierten Tonbandes', *Frequenz*, 1952, **6**, p. 319.

(9) Axon, P. E., Gilford, C. L. S., and Shorter, D. E. L.: 'Artificial Reverberation', *Proceedings I.E.E.*, Paper No. 1796 R, February, 1955 (**102** B, p. 624).

(10) Schrott, P.: 'Einfluss der Schrägstellung des Spaltes bei Intensitätsschrift', *Zeitschrift für Tecknische Physik*, 1936, **17**, p. 275.

(11) Wallace, R. L.: 'The Reproduction of Magnetically Recorded Signals', *Bell System Technical Journal*, 1951, **30**, p. 1146.

(12) Daniel, E. D., and Axon, P. E.: 'The Standardisation of Magnetic Tape Recording Systems', *B.B.C. Quarterly*, 1953–54, **8**, p. 235.

(13) Karlquist, O.: 'Calculation of the Magnetic Field in the Ferromagnetic Layer of a Magnetic Drum', *Transactions of the Royal Institute of Technology, Stockholm*, 1954, **86**, p. 3.

(14) Osmond, W. P.: 'The Effect of Particle Shape Variations on the Coercivity of Iron Oxide Powders', *Proceedings of the Physical Society*, B, 1954, **67**, p. 875.

(15) Muckenhirn, O. W.: 'Recording Demagnetisation in Magnetic Tape Recording', *Proceedings of the Institute of Radio Engineers*, 1951, **39**, p. 891.

(16) Westmijze, W. K.: 'The Principle of the Magnetic Recording and Reproduction of Sound', *Philips Technical Review*, 1953, **15**, p. 84.

(17) Axon, P. E., and Davies, H.: 'A Study of Frequency Fluctuations in Sound Recording and Reproducing Systems', *Proceedings I.E.E.*, Paper No. 799 R, January, 1949 (**96**, Part III, p. 65).

(18) Olson, H. F., Houghton, W. D., Morgan, A. R., Zenel, J., Artzt, M., Woodward, J. G., and Fischer, J. T.: 'A System for Recording and Reproducing Television Signals', *R.C.A. Review*, 1954, **15**, p. 3.

(19) Mullin, J. T.: 'Enregistrement magnétique des signaux video', *L'Onde Electrique*, 1954, **34**, p. 765.

(20) Werner, P. H.: 'Les propriétés mécaniques de diverses bandes magnetiques et leur repercussion sur la qualité de l'enregistrement', *Bulletin technique de l'administration des télégraphes et des téléphones suisses*, 1952, **30**, p. 173.

(21) Bozorth, R. M.: 'Ferromagnetism' (Van Nostrand Co.).

(22) Chynoweth, W. R.: 'Ferrite Heads for Recording in the Megacycle Range', *Tele-Tech.*, August, 1955.

(23) Kornei, O.: 'Magnetic Transducer Head for High Frequency Signals': U.S. Patent No. 2711945. 1955.

(24) Kornei, O.: 'Magnetic Transducer Head and Method of Making Same': U.S. Patent No. 721–755. 1951.

(25) Gouriet, G. G.: 'Spectrum Equalisation', *Wireless Engineer*, 1953, **30**, p. 112.

(26) Daniel, E. D., and Axon, P. E.: 'Accidental Printing in Magnetic Recording', *B.B.C. Quarterly*, 1950–51, **5**, p. 7.

(27) Kornei, O.: 'Survey of Flux-Responsive Magnetic Reproducing Heads', *Journal of the Acoustical Society of America*, 1955, **27**, p. 575.

(28) Daniel, E. D.: 'A Flux-Sensitive Reproducing Head for Magnetic Recording Systems', *Proceedings I.E.E.*, Paper No. 1856 R, July, 1955 (**102** B, p. 442).

621.372.2

Paper No. 2294 R
Mar. 1957

A THEORETICAL STUDY OF PROPAGATION ALONG TAPE LADDER LINES

By P. N. BUTCHER, Ph.D.

(*The paper was first received* 20*th June, and in revised form* 17*th October*, 1956.)

SUMMARY

Dispersion curves are calculated for single-ridge, double-ridge, single T-section and double T-section ladder lines in which the rungs of the ladder are thin tapes. Each structure is broken up into several regions having simple geometries. In each region the electromagnetic field is expanded as a series of suitable wave functions. Matching of the fields at the boundaries of the regions leads to an infinite set of homogeneous linear equations for the coefficients in the expansions. These equations have a non-trivial solution only if the determinant of their matrix is zero. The dispersion curves are obtained numerically from such determinantal equations. They confirm the qualitative predictions. Throughout the analysis, the ladder is approximated by a uniform sheet which conducts only in the direction of the tapes.

LIST OF SYMBOLS

A = Length of the tapes.
a = Total channel width.
A_{nr} = Coefficients in the equations relating U_r^e and U_r^0.
B = Distance of the tops of the ridges from the ladder.
b = Depth of the channels.
B_{nr} = Coefficients in the equations relating V_r^e and V_r^0.
C_{nr} = Coefficient which enters into the asymptotic form of the terms in the series for A_{nr} and B_{nr}.
D = Period of the ladder.
E = Electric vector.
E_p = Perturbed electric vector.
E_x, E_y, E_z = Cartesian components of the electric vector.
h = Distance of the top plate from the ladder in single-ridge and single-T structures.
H = Magnetic vector.
H_p = Perturbed magnetic vector.
H_x, H_y, H_z = Cartesian components of the magnetic vector.
J_{nm} = Elementary integral.
K_{nm} = Elementary integral.
k = Phase-change coefficient in free space.
$m \begin{cases} = \text{Even positive integer or zero in ridge structures.} \\ = \text{Odd positive integer in T structures.} \end{cases}$
n, r = Odd positive integers.
R = Region of free space in one period of a periodic guide.
δR = Region of free space removed from R by perturbation of the walls of the guide.
S = The part of the perfectly conducting surface of the perturbed guide which bounds δR.
t = Time.
U = E-wave generating function.
$U_r^0, U_r^e, U_m', U_r''$ = Coefficients in the expansions of U.
V = H-wave generating function.
$V_r^0, V_r^e, V_m', V_r''$ = Coefficients in the expansions of V.
x, y, z = Cartesian co-ordinates.
Z_0 = Intrinsic impedance of free space.
β = Axial phase-change coefficient of the fundamental space harmonic.
$\gamma_r = \sqrt{[\beta^2 - k^2 + (r\pi/A)^2]}$
$\gamma_m' = \sqrt{[\beta^2 - k^2 + (m\pi/a)^2]}$
δ_{nr} = Kronecker δ.
$\Delta_m \begin{cases} = 2, m = 0. \\ = 1, m \neq 0. \end{cases}$
ϵ_0 = Permittivity of free space.
λ = Free-space wavelength.
μ_0 = Permeability of free space.
ϕ = Wave function.
ω = Angular frequency.
ω_p = Perturbed angular frequency.

(1) INTRODUCTION

The strength of the interaction between the electron beam and the electromagnetic field in a travelling-wave tube is proportional to the fraction of the total electromagnetic energy which is stored in the region of the beam by the axial electric field of the space harmonic to which the beam is coupled. This provides one criterion by which to appraise the relative merits of different structures. Helices are good structures in this respect, but in the millimetre wavebands they are fragile and have low thermal capacities. These defects are avoided in ladder lines consisting of a periodic array of parallel straight conducting wires incorporated in a massive metal structure which short-circuits the array on each side. They can be made easily either by photo-etching a thin plate or by winding tape on to a rod with a suitable channel cut out of it.[1]

The dispersion curve of a ladder line depends on the cross-section of the wires and the shape of the surrounding structure. Leblond and Mourier[2] have considered the dispersion curves of lines in which the wires have a rectangular or circular cross-section and are thick enough in the direction normal to the plane of the ladder to ensure that each wire is shielded from all except its nearest neighbours. The present paper is concerned with the case in which the wires are thin tapes, so that there is very little shielding. Pierce[3] has given an elementary treatment which is applicable to lines of this type. The method used in the present paper provides an exact solution when the ladder is approximated by a uniform sheet which conducts only in the direction of the tapes. The dispersion curves are discussed qualitatively in the next Section before proceeding to the detailed analysis. The M.K.S. system of units is used throughout.

(2) QUALITATIVE TREATMENT OF THE DISPERSION CURVES

The structures to be discussed are shown in Fig. 1 and their dimensions are labelled there. It is convenient to use a right-handed system of rectangular Cartesian co-ordinate axes $0xyz$ with the z-axis parallel to the axis of the guide and $0xy$ located

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Butcher is now at the Stanford Electronics Laboratories, Stanford University, California, on leave from the Radar Research Establishment, Malvern.

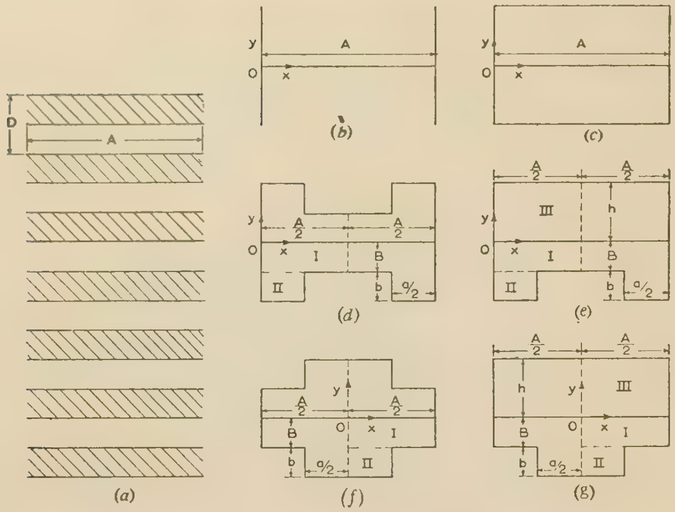


Fig. 1.—Tape ladder lines.

(*a*) Plan of the ladder. (*b*) Easitron structure. (*c*) Rectangular structure. (*d*) Double-ridge structure. (*e*) Single-ridge structure. (*f*) Double-T structure. (*g*) Single-T structure. The tapes are shaded in (*a*). (*b*) to (*g*) are transverse cross-sections, the ladder being in the plane $y = 0$ in every case.

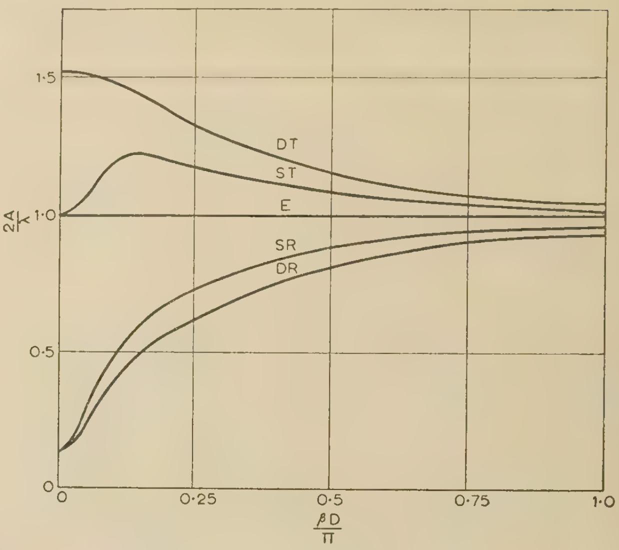


Fig. 2.—Sketch of the dispersion curves.

DT = Double-T structure.
ST = Single-T structure.
E = Easitron and rectangular structures.
SR = Single-ridge structure.
DR = Double-ridge structure.

and oriented as shown in the Figure. The simplest structure is that in which the array of tapes is short-circuited on either side by two vertical perfectly conducting side walls [Fig. 1(*b*)]. This will be called the 'easitron structure' because a ladder line of this type was used by Walker[4] in a distributed form of unloaded multi-resonator klystron called an 'easitron'. The array of tapes may be regarded as a multi-conductor transmission line. It can propagate a variety of TEM waves, each one corresponding to a different mode of excitation of the tapes. The relevant case is that in which there is simply a phase-change βD from one tape to the next, D being the period of the array and β the axial phase-change coefficient of the fundamental space harmonic (i.e. $|\beta D| < \pi$). Whatever the value of β, the short-circuited array can support TEM standing-waves at the frequency for which the free-space wavelength λ is twice the length of the tapes, i.e. at the first resonant frequency of the short-circuited tapes. The dispersion curve is the horizontal straight line labelled E in Fig. 2. There are, of course, higher pass bands, but they are of no interest in the present work.

There are several things to notice about the field distribution in the easitron structure. The field vectors are everwhere parallel to the side walls. The electric vector $\boldsymbol{E}$ is an even function of $x - A/2$, having maximum amplitude at $x = A/2$ and vanishing at the side walls. The magnetic vector $\boldsymbol{H}$ is an odd function of $x - A/2$, having maximum amplitude at the side walls and vanishing at $x = A/2$. Thus the vertical plane of symmetry $x = A/2$ is a 'magnetic wall'. From the symmetry of the structure it is clear that the z-component of the electric vector and the y-component of the magnetic vector are both even functions of y, while the y-component of the electric vector and the z-component of the magnetic vector are both odd functions of y. Both field vectors decrease in the y-direction more or less exponentially with an attenuation coefficient equal to β. This is an immediate consequence of Laplace's equation. The exact field distribution about a periodic array of parallel straight tapes is derived in another paper.[5]

The ladder can be completely enclosed in a rectangular guide by adding horizontal perfectly conducting plates on either side of it to bridge the side walls [Fig. 1(*c*)]. These two plates merely form two extra conductors in the multi-conductor transmission line considered above. The dispersion curve is absolutely unaltered, but the field distribution is modified when the horizontal plates are very close to the ladder. Moreover, the structure can now propagate the H_{01} mode of the empty rectangular guide, because this mode has no tangential electric field at the tapes and is not disturbed by them. This is a fast wave but it is of some interest in the subsequent discussion.

In both the easitron structure and the rectangular structure the slow wave has zero group velocity for all values of β. It can be made to propagate energy by suitably distorting the walls surrounding the ladder. Consider the rectangular structure—the easitron structure can be regarded as a limiting case. The frequency corresponding to any value of β can be reduced by pushing in a ridge at the centre of the top plate or the bottom plate, or both, where the magnetic field is weak and the electric field is strong [Figs. 1(*d*) and 1(*e*)]. This effect can be loosely ascribed to the increased capacitance between the centres of the horizontal plates and the ladder. It is better regarded as an immediate consequence of the theorem proved in the Appendix. When β is large, the field falls off rapidly in the y-direction and the ridges have only a small effect. In particular, the π-mode cut-off frequency is only just below the first resonant frequency of the short-circuited tapes. As β is reduced, the field spreads out further from the ladder and the frequency is depressed more by the ridges.

The zero-mode cut-off frequency can be approximated by considering the two uniform guides formed on either side of the ladder by covering it with a perfectly conducting plate. In general, the H_{01} mode of one of these guides will have a lower cut-off frequency than the H_{01} mode of the other. At cut-off there are no longitudinal currents in the walls of this guide and so the field and the frequency are largely unaffected by transverse slots. Thus we obtain the zero-mode field and frequency of the slow wave in the ridge ladder line. By considering the H_{01} mode of the other uniform guide we obtain the zero-mode field and frequency of the fast wave which is a perturbation of the H_{01} mode of the rectangular ladder line. When the structure is symmetrical about the plane of the ladder [Fig. 1(*d*)], the fast wave is simply the H_{01} mode of the empty guide, which is undisturbed

by the tapes. It has approximately the same zero-mode cut-off frequency as the slow wave. The cut-off frequencies are not altered very much if the tapes in the ladder are replaced by wires with some other cross-section, provided that their cross-sectional dimensions are small.

To be definite, we shall confine our attention to the double-ridge structure shown in Fig. 1(*d*) and the corresponding single-ridge structure shown in Fig. 1(*e*). The former is symmetrical about the plane of the ladder, the latter is not. Both structures are symmetrical about the plane $x = A/2$. For large values of β, the depression of the frequency due to two ridges will be roughly twice that due to one. However, the zero-mode cut-off frequency is roughly the same in both cases, being approximately equal to that of the H_{01} mode in the uniform ridge guide formed below the ladder by covering it with a perfectly conducting plate. The dispersion curves therefore have the forms labelled DR (double ridge) and SR (single ridge) in Fig. 2.

The slow wave can also be made to propagate energy by pushing in ridges at the sides of the top and bottom plates of the guide where the electric field is weak and the magnetic field is strong. The frequency corresponding to any value of β is then increased. This effect can be loosely ascribed to the decrease of the inductance at the sides of the guide, but it is better regarded in the light of the theorem proved in the Appendix. The discussion of the dispersion curves and the cut-off frequencies is similar to that given above. To be definite, we shall consider only the double-T structure shown in Fig. 1(*f*) and the corresponding single-T structure shown in Fig. 1(*g*). (Notice that the origin of co-ordinates has been moved from the side wall in these Figures. The reason for this will be made clear in the next Section.) The double-T structure is symmetrical about the planes $x = 0$ and $y = 0$; the single-T structure is symmetrical only about the plane $x = 0$. The dispersion curves are those labelled DT (double T) and ST (single T) in Fig. 2. The peculiar behaviour of the latter curve comes about because the zero-mode cut-off frequency is approximately equal to that of the H_{01} mode of the uniform rectangular guide formed above the ladder when it is covered with a perfectly conducting plate, and this is the same as the first resonant frequency of the short-circuited tapes. The cut-off frequency of the H_{01} mode of the uniform T-section guide formed below the ladder is higher and applies to the fast wave. Notice that the fast wave cannot be propagated through either structure at frequencies within the pass band of the slow wave. However, in the single-T structure there are two distinct 'slow waves' (one of which has in fact a high phase velocity) which can be propagated in either direction through the guide.

The π-mode cut-off frequency of any of these ladder lines can be raised by using shorter tapes running between horizontal plates supported by the side walls. The π-mode cut-off frequency is still approximately equal to the first resonant frequency of the short-circuited tapes, and is thus inversely proportional to their length. This is a valuable technique for broadening the pass band of ridge structures. The slow wave can be made to propagate energy through a rectangular structure in this way, without introducing any ridges. However, short tapes will not be considered further in the paper because the field-matching technique which will be employed in the subsequent analysis becomes more complicated in this case (although it is still applicable).

(3) THE DISPERSION EQUATIONS

The double-ridge and single-ridge structures will be considered in detail. The trivial changes that are necessary in the analysis of the double-T and single-T structures will be indicated later on. The electromagnetic field is a superposition of space harmonics which are coupled together by the periodic boundary conditions at the ladder. However, by making two reasonable assumptions about the field at the ladder, it is possible to find approximate boundary conditions which involve only the fundamental space harmonic. If E_x is negligible in the gaps between the tapes, then the fundamental space harmonic of E_x is zero on both sides of the plane of the ladder. If H_x is negligible on the tapes (because the current in the tapes flows mostly in the x-direction), then the fundamental space harmonic of H_x is continuous across the plane of the ladder. Finally, without any approximation, the fundamental space harmonic of E_z is continuous across the plane of the ladder. These boundary conditions, together with those at the perfectly conducting walls, are sufficient to characterize completely the fundamental space harmonic of the electromagnetic field. They allow the calculation of an approximate dispersion equation. The approximations made are justified when the tape width and the gap width are both small. Indeed, as these dimensions tend to zero, the ladder simulates a smooth sheet which conducts in the x-direction but not in the z-direction, and the above boundary conditions are rigorously correct.

The complex electric and magnetic vectors of the fundamental space harmonic contain the common factor $\exp j(\omega t - \beta z)$, where ω is the angular frequency and t is the time. This factor will be omitted from the following analysis. The field vectors are then functions of x and y only. They can be derived from an E-wave generating function U and an H-wave generating function V through the equations:[6]

$$\left.\begin{aligned} E_x &= -j\left(\beta\frac{\partial U}{\partial x} + k\frac{\partial V}{\partial y}\right) \\ E_y &= -j\left(\beta\frac{\partial U}{\partial y} - k\frac{\partial V}{\partial x}\right) \\ E_z &= (k^2 - \beta^2)U \end{aligned}\right\} \quad . \quad . \quad . \quad (1a)$$

$$\left.\begin{aligned} Z_0H_x &= -j\left(\beta\frac{\partial V}{\partial x} - k\frac{\partial U}{\partial y}\right) \\ Z_0H_y &= -j\left(\beta\frac{\partial V}{\partial y} + k\frac{\partial U}{\partial x}\right) \\ Z_0H_z &= (k^2 - \beta^2)V \end{aligned}\right\} \quad . \quad . \quad . \quad (1b)$$

where Z_0 is the intrinsic impedance of free-space and k is the free-space phase-change coefficient $2\pi/\lambda$. The generating functions satisfy the two-dimensional wave equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + (k^2 - \beta^2)\phi = 0 \quad . \quad . \quad . \quad (2)$$

It follows immediately, from eqns. (1) and the boundary conditions at a perfectly conducting surface, that U and $\partial V/\partial x$ must vanish on the vertical conducting walls, and U and $\partial V/\partial y$ must vanish on the horizontal conducting walls. Moreover, as was found in the easitron structure, the symmetry about the plane $x = A/2$ makes this plane a magnetic wall on which the tangential magnetic field vanishes. Hence V and $\partial U/\partial x$ must vanish there. In each of the regions of the double-ridge and single-ridge structures indicated in Figs. 1(*d*) and 1(*e*), i.e. the regions I, II and III, the generating functions can be expanded as well-behaved series of the separable wave functions which satisfy these boundary conditions (except those relating to the top of the ridge). Thus, in region I,

$$U = \sum_r (U_r^e \cosh\gamma_r y + U_r^0 \sinh\gamma_r y)\sin\frac{r\pi x}{A} \quad . \quad (3a)$$

$$V = \sum_r (V_r^e \cosh\gamma_r y + V_r^0 \sinh\gamma_r y)\cos\frac{r\pi x}{A} \quad . \quad (3b)$$

where $$\gamma_r = \sqrt{[\beta^2 - k^2 + (r\pi/A)^2]} \quad . \quad . \quad . \quad (3c)$$

and r takes all positive odd integral values. In region II,

$$U = \sum_m U'_m \sinh \gamma'_m(y + B + b) \sin \frac{m\pi x}{a} \quad . \quad . \quad (4a)$$

$$V = \sum_m V'_m \cosh \gamma'_m(y + B + b) \cos \frac{m\pi x}{a} \quad . \quad . \quad (4b)$$

where $$\gamma'_m = \sqrt{[\beta^2 - k^2 + (m\pi/a)^2]} \quad . \quad . \quad . \quad (4c)$$

and m takes all positive even integral values and the value zero. The coefficients U^e_r, U^0_r, V^e_r, V^0_r, U'_m and V'_m remain to be determined.

The boundary conditions at the plane $y = -B$ are that E_x and E_z must vanish on the ridge and must match at the join of I and II together with H_x and H_z. This implies that (i) U must vanish on the ridge and match at the join, (ii) $\partial V/\partial y$ must vanish on the ridge and match at the join, (iii) V must match at the join, and (iv) $\partial U/\partial y$ must match at the join. It is a simple matter to give explicit expression to these boundary conditions in terms of the expansions (3) and (4). Multiplying the first by $\sin n\pi x/A$ and integrating (with respect to x) from 0 to $A/2$, multiplying the second by $\cos n\pi x/A$ and integrating from 0 to $A/2$, multiplying the third by $\cos m\pi x/a$ and integrating from 0 to $a/2$, and multiplying the fourth by $\sin m\pi x/a$ and integrating from 0 to $a/2$ we obtain four sets of equations relating the coefficients in the expansions of U and V:

$$U^e_n \cosh \gamma_n B - U^0_n \sinh \gamma_n B = \sum_m U'_m \mathrm{J}_{nm} \sinh \gamma'_m b \quad (5a)$$

$$\gamma_n(-V^e_n \sinh \gamma_n B + V^0_n \cosh \gamma_n B) = \sum_m V'_m \mathrm{K}_{nm} \gamma'_m \sinh \gamma'_m b \quad (5b)$$

$$\Delta_m V'_m \cosh \gamma'_m b = \frac{A}{a} \sum_r (V^e_r \cosh \gamma_r B - V^0_r \sinh \gamma_r B)\mathrm{K}_{rm} \quad (5c)$$

$$U'_m \gamma'_m \cosh \gamma'_m b = \frac{A}{a} \sum_r \gamma_r(-U^e_r \sinh \gamma_r B + U^0_r \cosh \gamma_r B)\mathrm{J}_{rm} \quad (5d)$$

where n takes all positive odd integral values,

$$\mathrm{J}_{nm} = \frac{4}{A} \int_0^{a/2} dx \sin \frac{n\pi x}{A} \sin \frac{m\pi x}{a} \quad . \quad . \quad . \quad (6a)$$

$$\mathrm{K}_{nm} = \frac{4}{A} \int_0^{a/2} dx \cos \frac{n\pi x}{A} \cos \frac{m\pi x}{a} \quad . \quad . \quad . \quad (6b)$$

and $$\Delta_m = 2, m = 0$$
$$= 1, m \neq 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

which takes care of the unfortunate fact that the integral of $\cos^2 m\pi x/a$ from 0 to $a/2$ is equal to $a/2$ when $m = 0$, and is equal to $a/4$ when $m \neq 0$. Substituting U'_m and V'_m from eqns. (5c) and (5d) into eqns. (5a) and (5b) and interchanging the order of summation, we obtain

$$U^e_n \cosh \gamma_n B - U^0_n \sinh \gamma_n B = \sum_r A_{nr}(-U^e_r \sinh \gamma_r B + U^0_r \cosh \gamma_r B) \quad (8a)$$

$$-V^e_n \sinh \gamma_n B + V^0_n \cosh \gamma_n B = \sum_r B_{nr}(V^e_r \cosh \gamma_r B - V^0_r \sinh \gamma_r B) \quad (8b)$$

where $$A_{nr} = \left(\frac{A}{a}\right)\gamma_r \sum_m \mathrm{J}_{nm}\mathrm{J}_{rm} \tanh \gamma'_m b/\gamma'_m \quad . \quad . \quad (9a)$$

$$B_{nr} = \left(\frac{A}{a}\right)\frac{1}{\gamma_n} \sum_m \mathrm{K}_{nm}\mathrm{K}_{rm}\gamma'_m \tanh \gamma'_m b/\Delta_m \quad . \quad . \quad (9b)$$

Eqns. (8) can be written in the more convenient form

$$\sum_r [U^e_r(A_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) - U^0_r(A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B)] = 0 \quad (10a)$$

$$\sum_r [V^0_r(B_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) - V^e_r(B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B)] = 0 \quad (10b)$$

where $$\delta_{nr} = 1, n = r$$
$$= 0, n \neq r \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

is the Kronecker δ.

So far, the analysis has been common to both double-ridge and single-ridge structures. The matching of the fields at the plane of the ladder, $y = 0$, is very simple because it can be done term by term, but it is different in the two cases. Consider the double-ridge structure. The symmetry of the field about the plane of the ladder is the same as was found in the easitron structure. In particular, E_z is an even function of y, and H_z is an odd function of y. Hence, from eqns. (1), U and V must be even and odd functions of y, respectively. It follows that E_z is necessarily continuous at $y = 0$ while H_x will be continuous at $y = 0$ if it vanishes there. The boundary conditions at the plane of the ladder therefore reduce to the requirement that both E_x and H_x should vanish there. Substituting the series for U and V given in eqns. (3) into eqns. (1), setting $y = 0$, and equating to zero term by term the resulting series for E_x and H_x, it is found that

$$\beta\left(\frac{r\pi}{A}\right)U^e_r + k\gamma_r V^0_r = 0 \quad . \quad . \quad . \quad (12a)$$

$$\beta\left(\frac{r\pi}{A}\right)V^e_r + k\gamma_r U^0_r = 0 \quad . \quad . \quad . \quad (12b)$$

Solving for U^e_r and V^e_r and substituting in eqns. (10) gives

$$\sum_r \left[V^0_r \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(A_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + U^0_r(A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \right] = 0 \quad (13a)$$

$$\sum_r \left[V^0_r(B_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + U^0_r \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \right] = 0 \quad (13b)$$

This infinite set of homogeneous linear equations for U^0_r and V^0_r has a non-trivial solution only if the infinite determinant of its matrix vanishes, i.e. only if

$$\begin{vmatrix} \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(A_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B), & A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B \\ B_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B, & \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \end{vmatrix} = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

where it is understood that the array is to be expanded in blocks of 2×2 elements the nrth of which is given above. Notice that r varies horizontally and n varies vertically, both taking only odd positive integral values. Eqn. (14) is the dispersion equation for the double-ridge structure.

Turning now to the single-ridge structure, it is necessary to consider region III [Fig. 1(*e*)]. Here U and V can be expanded as follows:

$$U = \sum_r U_r'' \sinh \gamma_r(y - h) \sin \frac{r\pi x}{A} \quad . \quad . \quad (15a)$$

$$V = \sum_r V_r'' \cosh \gamma_r(y - h) \cos \frac{r\pi x}{A} \quad . \quad . \quad (15b)$$

where r takes only positive odd integral values. Substituting U and V from eqns. (3) and (15) into eqns. (1), equating to zero term by term the series for E_x on either side of the ladder, and matching term by term the series for E_z and H_x on either side of the ladder, it is found that

$$\beta\left(\frac{r\pi}{A}\right) U_r^e + k\gamma_r V_r^0 = 0 \quad . \quad . \quad . \quad . \quad (16a)$$

$$\beta\left(\frac{r\pi}{A}\right) U_r'' + k\gamma_r V_r'' = 0 \quad . \quad . \quad . \quad . \quad (16b)$$

$$U_r^e = - U_r'' \sinh \gamma_r h \quad . \quad . \quad . \quad . \quad (16c)$$

$$\beta\left(\frac{r\pi}{A}\right) V_r^e + k\gamma_r U_r^0 = \left[\beta\left(\frac{r\pi}{A}\right) V_r'' + k\gamma_r U_r''\right] \cosh \gamma_r h \quad (16d)$$

Solving for U_r^e and V_r^e we have, after a little manipulation,

$$\left.\begin{aligned} U_r^e &= -\frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right) V_r^0 \\ V_r^e &= -\frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right) U_r^0 - \coth \gamma_r h\left[1 - \left(\frac{k}{\beta}\right)^2\left(\frac{A\gamma_r}{r\pi}\right)^2\right] V_r^0 \end{aligned}\right\} \quad (17)$$

Substitution for U_r^e and V_r^e in eqns. (10) gives an infinite set of homogeneous linear equations for U_r^0 and V_r^0:

$$\sum_r \left[V_r^0 \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(A_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + U_r^0 (A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \right] = 0 \quad . \quad . \quad . \quad . \quad (18a)$$

$$\sum_r \left(V_r^0 \left\{ (B_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + \coth \gamma_r h \left[1 - \left(\frac{k}{\beta}\right)^2 \left(\frac{\gamma_r A}{r\pi}\right)^2\right] (B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \right\} + U_r^0 \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \right) = 0 \quad . \quad . \quad (18b)$$

The determinantal equation can be written down by inspection. It can easily be cast into the more symmetrical form

$$\begin{vmatrix} \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)[(A_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + \coth \gamma_r h (A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B)], & A_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B \\ [(B_{nr} \sinh \gamma_r B + \delta_{nr} \cosh \gamma_r B) + \coth \gamma_r h (B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B)], & \frac{k}{\beta}\left(\frac{\gamma_r A}{r\pi}\right)(B_{nr} \cosh \gamma_r B + \delta_{nr} \sinh \gamma_r B) \end{vmatrix} = 0 \quad (19)$$

The remarks following eqn. (14) also apply here. Eqn. (19) is the dispersion equation of the single-ridge structure.

The analysis of the double-T and single-T structures is only trivially different, particularly if the origin of co-ordinates is moved on to the vertical plane of symmetry and the regions I, II and III are defined as indicated in Figs, 1(*f*) and 1(*g*). The plane $x = 0$ is then a magnetic wall and the plane $x = A/2$ is perfectly conducting; it was the other way round in the ridge structures. Hence sines are replaced by cosines in the above analysis and vice versa. Moreover, because there is still a perfectly conducting wall on the right-hand side of region II, the summation variable m, which appears in the expansions of U and V in this region [eqns. (4) with sines and cosines interchanged] and elsewhere in the analysis, now takes odd positive integral values instead of even ones and it no longer takes the value zero. The dispersion equations for the double-T and single-T structures are given formally by eqns. (14) and (19), respectively, but m takes only odd positive integral values in eqns. (9), and sines and cosines must be interchanged in eqns. (6).

(4) THE NUMERICAL RESULTS

The apparently formidable determinantal equations which were obtained in the previous Section are surprisingly easy to handle numerically. It is convenient to regard $\beta^2 - k^2$ as the independent variable when carrying out the calculations. Once this is fixed, together with the dimensions of the structure, everything in the determinantal equation can be evaluated except the ratio k/β, which occurs in various elements of the determinant. A $2n \times 2n$ approximation to the determinant (obtained by taking only the first $2n$ rows and the first $2n$ columns) then provides an nth-order polynomial in $(k/\beta)^2$ whose lowest positive zero is the '$2n \times 2n$ approximation' to the value of $(k/\beta)^2$ appropriate to the slow wave. The calculation of k and β from $\beta^2 - k^2$ and $(k/\beta)^2$ is trivial.

It is perhaps worth while to mention the method of calculation of the series for A_{nr} and B_{nr} which are given in eqns. (9). The integrals J_{nm} and K_{nm} which are defined by eqns. (6) can be evaluated immediately. When m is large they decrease as m^{-1} and m^{-2}, respectively (this is still true in the case of the T-structures). Hence, the terms of the series A_{nr} and B_{nr} tend asymptotically to the form $C_{nr}m^{-3}$, where C_{nr} is easily evaluated and is independent of m. The series can therefore be summed quickly by subtracting $\sum_m C_{nr}m^{-3}$ term by term, summing the rapidly convergent series formed thereby to an adequate number of terms, and finally adding on $C_{nr} \sum_m m^{-3}$. The slowly convergent series $\sum_m m^{-3}$ can be evaluated once and for all by means of the Euler–Maclaurin formula.[7]

The convergence of the successive approximations is extremely rapid. The 2×2 and 4×4 approximations differ most when the ridges are very close to the ladder, but the difference is still very small for the least value of B/A which was considered in each case. Thus, for the double-ridge, single-ridge, double-T and single-T structures with $B/A = 0{\cdot}05$, $0{\cdot}05$, $0{\cdot}04$ and $0{\cdot}02$, respectively, the two approximations to k differ by less than 1% when βB is greater than $0{\cdot}21$, $0{\cdot}16$, $0{\cdot}17$ and $0{\cdot}16$, respectively, and the maximum difference calculated was $3{\cdot}3\%$, $3{\cdot}2\%$, $4{\cdot}4\%$ and $2{\cdot}6\%$, respectively. The 2×2 approximations to the dis-

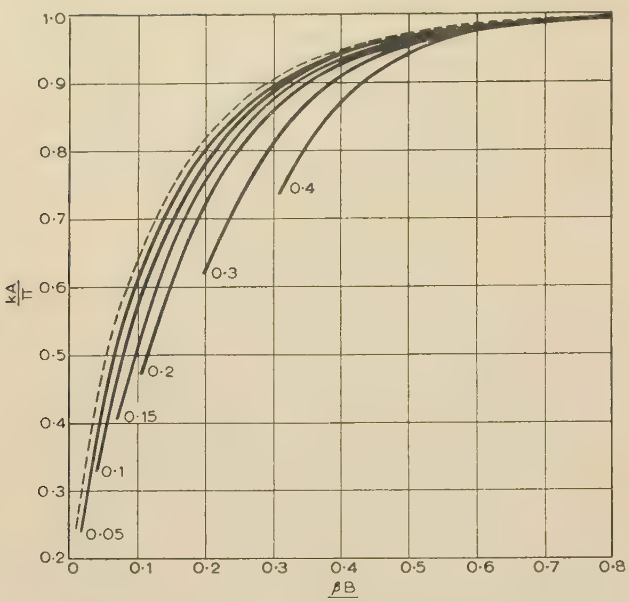


Fig. 3.—Dispersion curves for double-ridge structures.

4 × 4 approximation, $a/A = 0{\cdot}5$, $b = \infty$, B/A as indicated on the curves. The dashed curve is derived from an analysis similar to Pierce's.

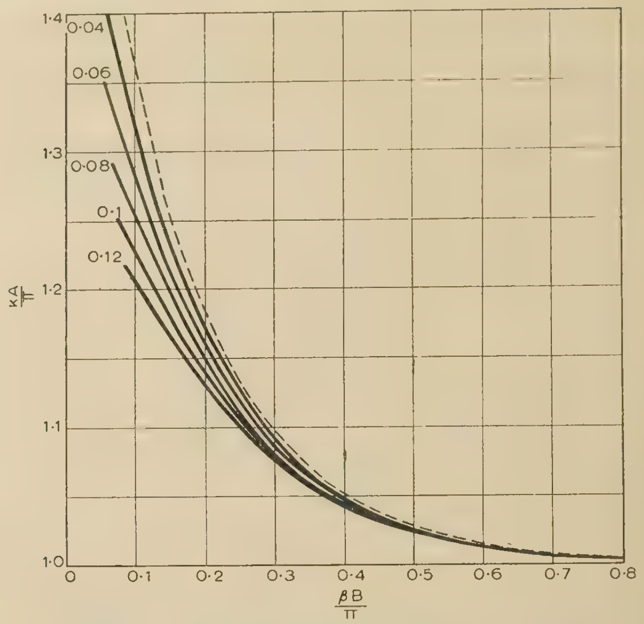


Fig. 5.—Dispersion curves for double-T structures.

4 × 4 approximation, $a/A = 0{\cdot}5$, $b = \infty$, B/A as indicated on the curves. The dashed curve is taken from Pierce's paper.

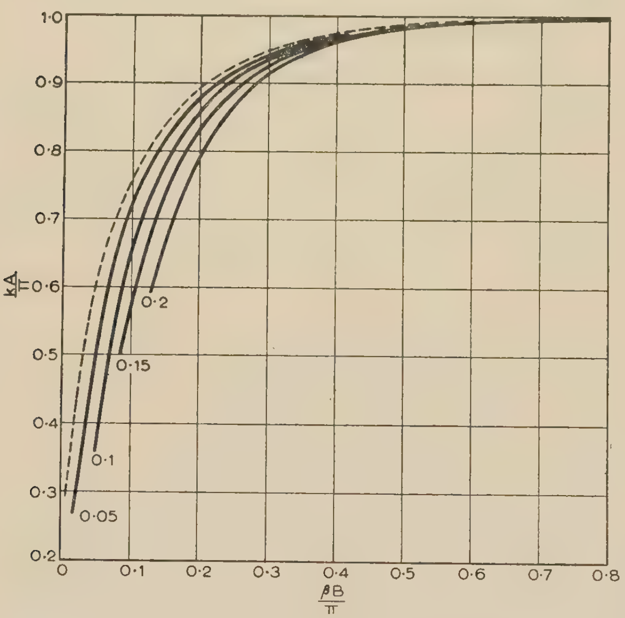


Fig. 4.—Dispersion curves for single-ridge structures.

4 × 4 approximation, $a/A = 0{\cdot}5$, $b = \infty$, $h = \infty$, B/A as indicated on the curves. The dashed curve is taken from Pierce's paper.

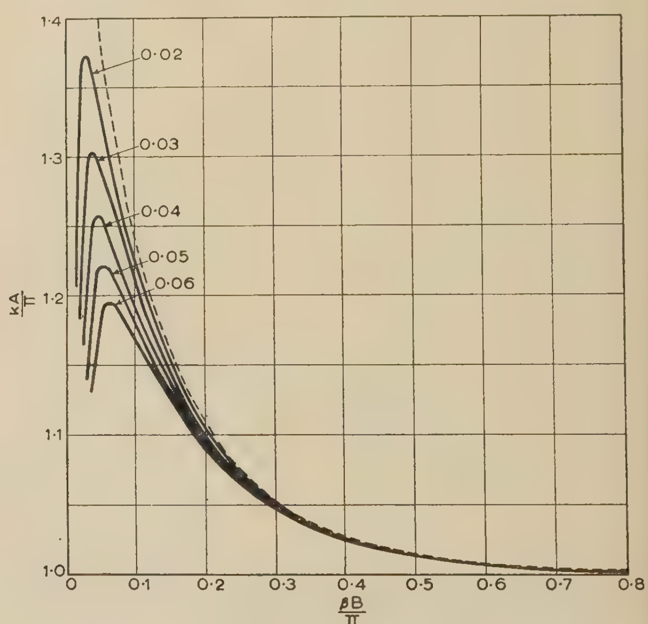


Fig. 6.—Dispersion curves for single-T structures.

4 × 4 approximation, $a/A = 0{\cdot}5$, $b = \infty$, $h = \infty$, B/A as indicated on the curves. The dashed curve is derived from an analysis similar to Pierce's.

persion curves are adequate for most purposes even when the ridges are very close to the ladder. The 4 × 4 approximations are a little more accurate; they are shown in Figs. 3–6 for $a/A = 0{\cdot}5$, $b = \infty$, $h = \infty$ and various values of B/A. It is unlikely that the dispersion curves would be altered significantly by taking higher-order approximations to the determinants in eqns. (14) and (19).

It will be noticed that, for given values of B/A and βB, the magnitude of $1 - kA/\pi$ is larger in the ridge structures than it is in the corresponding T-structures. This is because the term with $m = 0$ is missing from the expansions of U and V in region II of the T-structures. This term is relatively slowly attenuated in the y-direction, and its removal makes the distortion of the structure less effective. However, for travelling-wave tube applications, it is possible to use smaller values of B/A in T-structures than in ridge structures, provided that the electron beam is confined to the centre channel (in ridge structures the minimum value of B which can be used is determined by the beam cross-section).

The behaviour of the dispersion curves for small values of β requires comment. The zero-mode cut-off frequency of the double-ridge and single-ridge structures is zero when $b = \infty$.

The dispersion curves approach the line $k = \beta$ from the right as β tends to zero. They cannot cross over this line because it is the boundary of a forbidden region[8] in which γ_0 is imaginary and the terms with $m = 0$ in the expansions (4) are radiative. The corresponding expansions for the T-structures are obtained by interchanging sines and cosines and giving m positive odd integral values. There are no terms with $m = 0$. The terms with $m = 1$ are radiative to the left of the line $k^2 - \beta^2 = (\pi/a)^2$, but this lies outside the range of the calculations. In the single-T structure with $h = \infty$, however, the terms with $r = 1$ in the expansions corresponding to expansions (15), i.e. with sines and cosines interchanged, become radiative to the left of the line $k^2 - \beta^2 = (\pi/A)^2$. The dispersion curves keep to the right of this line.

The dashed curves plotted in Figs. 3 to 6 are taken from Pierce's paper[3] in the case of the single-ridge and double-T structures, and are derived from a similar analysis in the case of the double-ridge and single-T structures. They are independent of B/A and agree best with the solid curves for small values of B/A. This is what one would expect. Pierce approximates to the field distribution by taking TEM standing waves on either side of the plane $x = a/2$. He matches the fields at this plane without taking account of the lumped capacitance there. The error due to this omission will be smaller at low frequencies, i.e. when A is large.

The dispersion curves of structures in which b and h are finite differ from those plotted in Figs. 3 to 6 (see Section 2). However, the difference is small, provided that β is sufficiently large to make the hyperbolic tangents and cotangents in eqns. (9), (14) and (19) very close to unity. The numerical results are not difficult to obtain when b and h are finite, but it would be tedious to review them here.

It only remains to consider the effect of varying the total channel width, a. In Pierce's analysis, when β, A, B and b are held constant, the deviation of the frequency from the first resonant frequency of the short-circuited tapes is a maximum when $a/A = 0{\cdot}5$. This is not so in the present treatment. Consider the double-ridge structure. Solving the 2×2 approximation to eqn. (14) for $(k/\beta)^2$ gives immediately

$$\left(\frac{k}{\beta}\right)^2 = \left(\frac{\pi}{A\gamma_1}\right)^2 \frac{(A_{11} + \tanh \gamma_1 B)(B_{11} + \coth \gamma_1 B)}{(A_{11} + \coth \gamma_1 B)(B_{11} + \tanh \gamma_1 B)} \qquad (20)$$

When β is very large, $k \sim \pi/A$, $\gamma_1 \sim \beta$ and eqn. (20) takes the approximate form

$$1 - (kA/\pi) \sim \frac{2(B_{11} - A_{11}) \exp(-2\beta B)}{(B_{11} + 1)(A_{11} + 1)} \quad . \quad . \quad (21)$$

Now, from eqns. (6) and (9), when β is very large,

$$A_{11} \sim \frac{A}{a} \sum_m \left(\frac{4}{A}\int_0^{a/2} dx \sin\frac{\pi x}{A} \sin\frac{m\pi x}{a}\right)^2$$

$$= \frac{4}{A}\int_0^{a/2} dx \sin\frac{\pi x}{A} \sum_m \sin\frac{m\pi x}{a} \times \frac{4}{a}\int_0^{a/2} dx' \sin\frac{\pi x'}{A} \sin\frac{m\pi x'}{a}$$

$$= \frac{4}{A}\int_0^{a/2} dx \sin^2\frac{\pi x}{A} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

since the series is simply a Fourier expansion of $\sin \pi x/A$ on the union of regions I and II. Similarly

$$B_{11} \sim \frac{4}{A}\int_0^{a/2} dx \cos^2\frac{\pi x}{A} \quad . \quad . \quad . \quad . \quad (23)$$

Substituting these asymptotic values of A_{11} and B_{11} into the right-hand side of eqn. (21) gives finally

$$1 - (kA/\pi) \sim \frac{4\pi \exp(-2\beta B) \sin\dfrac{\pi a}{A}}{\pi^2\left(1 + \dfrac{a}{A}\right)^2 - \sin^2\dfrac{\pi a}{A}} \quad . \quad . \quad (24)$$

for the double-ridge structure when β is very large. For the single-ridge structure with $h = \infty$ (the only case considered here), the asymptotic value of $1 - kA/\pi$ is half of this. For the T-structures there is merely a change of sign. The optimum value of a/A will be defined as that which maximizes the asymptotic value of $|1 - kA/\pi|$ when β and B are held fixed. From eqn. (24) it is found that, in the 2×2 approximation, the optimum value is $a/A = 0{\cdot}366$ for all the structures considered here. However, the right-hand side of eqn. (24) is a slowly varying function of a/A near the optimum and has fallen by only 9% from its maximum value when $a/A = 0{\cdot}25$ or $0{\cdot}5$. Thus the value of a/A is not critical when β is very large; nor is it critical for smaller values of β. To illustrate this, the 2×2 approximations to the dispersion curves of a double-ridge structure, a single-ridge structure,

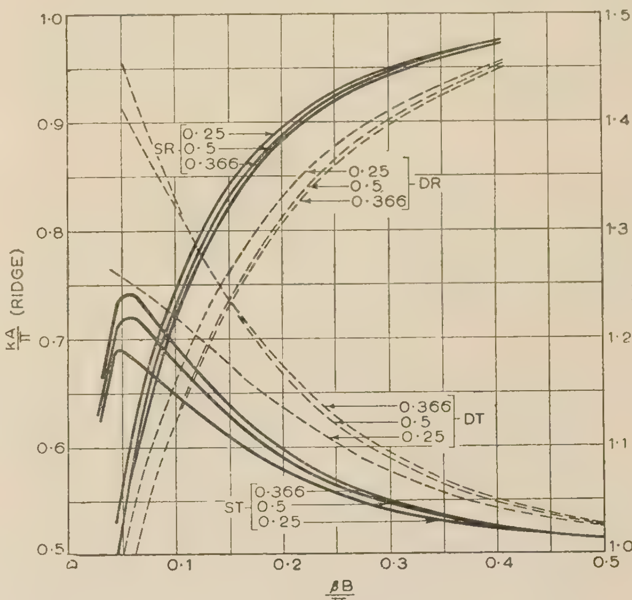


Fig. 7.—Dependence of the dispersion curves on the total channel width.

2×2 approximation, $B/A = 0{\cdot}05$, $b = \infty$, $h = \infty$, a/A as indicated on the curves. The scale on the left applies to single ridge (SR) and double (DR) structures. The scale on the right applies to single-T (ST) and double-T (DT) structures. The curves for double-ridge and double-T structures are dashed in order to make the diagram clearer.

a double-T structure and a single-T structure are drawn in Fig. 7 for the cases $a/A = 0{\cdot}25$, $0{\cdot}366$ and $0{\cdot}5$, when $b = \infty$, $h = \infty$ and $B/A = 0{\cdot}05$. The dispersion curve of the double-T structure is the most sensitive to variation of a/A.

(5) CONCLUSION

A method has been given for calculating the dispersion curves of the tape ladder lines without too much laborious computation. In every case, the dispersion curve for large values of the axial phase-change coefficient is primarily determined by the length of the tapes and the distance of the tops of the ridges from them. It is insensitive to the position of the bottoms of the channels (and the position of the top plate in single-ridge and single-T structures), provided that they are kept in weak fields. It is also

insensitive to changes of the total channel width, provided that this is kept within about 30% of the optimum value, which is estimated as 0·366 times the length of tapes. The width of the tapes and the period of the ladder do not enter into the analysis of the dispersion equations. Provided that these lengths are small enough to allow the application of the approximate boundary conditions at the ladder, their precise values are of significance only in so far as the period of the ladder determines the largest value of β which it is necessary to consider, i.e. π/D.

Single-ridge structures have already been used in millimetre-waveband backward-wave oscillators.[1,9] The fundamental space harmonic is a forward wave. Backward-wave interaction is obtained by synchronizing the electrons with the first reverse space harmonic. In T-structures, on the other hand, the fundamental space harmonic is a backward wave. The electrons could be directly coupled to this in a backward-wave tube. Any of the structures could, of course, be used in forward-wave tubes by coupling the electrons to an appropriate space harmonic. However, this is not the place to discuss their possible applications in detail.

Three outstanding problems remain: the estimation of the coupling impedance, the estimation of the error due to the simplification of the boundary conditions at the ladder, and the treatment of structures in which the matching of the fields on either side of the ladder cannot be done term by term. The first of these problems will be considered in the succeeding paper.[5]

(6) ACKNOWLEDGMENTS

The author is indebted to Heather M. Golby, who carried out the computations on the T.R.E. computer. He would also like to thank H. W. Duckworth, J. S. Thorp and R. J. Armstrong, of the Radar Research Establishment, and J. H. Collins and F. Gill, of The General Electric Co. Ltd., for many stimulating discussions.

(7) REFERENCES

(1) Karp, A.: 'Traveling-Wave Tube Experiments at Millimetre Wavelengths with a New, Easily Built, Space Harmonic Circuit', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 41.

(2) Leblond, A., and Mourier, G.: 'Étude des lignes à barreaux à structure périodique', *Annales de Radioélectricité*, 1954, **9**, p. 180, Part 2, p. 311.

(3) Pierce, J. R.: 'Propagation in Linear Arrays of Parallel Wires', *Transactions of the Institute of Radio Engineers*, 1955, **ED-2**, No. 1, p. 13.

(4) Walker, L. R.: Bell Telephone Laboratories, U.S.A., unpublished.

(5) Butcher, P. N.: 'The Coupling Impedance of Tape Structures' (see next page).

(6) Schelkunoff, S. A.: 'Electromagnetic Waves' (D. Van Nostrand, New York, 1943), Chapter 10.

(7) Jeffreys, H., and Jeffreys, B. S.: 'Methods of Mathematical Physics' (Cambridge University Press, 1946), p. 254.

(8) Sensiper, S.: 'Electromagnetic Wave Propagation on Helical Structures', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 149.

(9) Duckworth, H. W., and Thorp, J. S.: Radar Research Establishment, England, unpublished.

(10) Slater, J. C.: 'Microwave Electronics' (D. Van Nostrand, New York, 1950), Chapter 4.

(11) Casimir, H. B. G.: 'On the Theory of Electromagnetic Waves in Resonant Cavities', *Philips Research Reports*, 1951, **6**, p. 162.

(8) APPENDIX: PERTURBATION OF THE WALLS OF A PERIODIC WAVEGUIDE

Consider a perfectly conducting periodic waveguide. Let R be the region of free space bounded by the perfectly conducting surfaces and the transverse planes $z = 0$ and $z = D$. Within R the complex electric vector $\boldsymbol{E}$ and the complex magnetic vector $\boldsymbol{H}$ satisfy Maxwell's equations

$$\operatorname{curl} \boldsymbol{E} + j\omega\mu_0\boldsymbol{H} = 0 \quad . \quad . \quad . \quad . \quad . \quad (25)$$

$$\operatorname{curl} \boldsymbol{H} - j\omega\epsilon_0\boldsymbol{E} = 0 \quad . \quad . \quad . \quad . \quad . \quad (26)$$

where ω is the angular frequency considered and ϵ_0 and μ_0 are, respectively, the permittivity and permeability of free space. Suppose that the walls of the structure are pushed in so that a region δR is removed from R. The perturbed complex electric vector $\boldsymbol{E}_p$ and the perturbed magnetic vector $\boldsymbol{H}_p$ satisfy Maxwell's equations

$$\operatorname{curl} \boldsymbol{E}_p + j\omega_p\mu_0\boldsymbol{H}_p = 0 \quad . \quad . \quad . \quad . \quad (27)$$

$$\operatorname{curl} \boldsymbol{H}_p - j\omega_p\epsilon_0\boldsymbol{E}_p = 0 \quad . \quad . \quad . \quad . \quad (28)$$

where ω_p is the perturbed frequency. Taking the scalar product of eqns. (25) and (26), and the complex conjugates of eqns. (27) and (28) with $-\boldsymbol{H}_p^*$, $\boldsymbol{E}_p^*$, $-\boldsymbol{H}$ and $\boldsymbol{E}$, respectively, and adding, it is found that

$$j(\omega_p - \omega)(\mu_0\boldsymbol{H}\cdot\boldsymbol{H}_p^*) + \epsilon_0\boldsymbol{E}\cdot\boldsymbol{E}_p^* + \operatorname{div}(\boldsymbol{H}_p^* \times \boldsymbol{E} + \boldsymbol{H} \times \boldsymbol{E}_p^*) = 0 \quad (29)$$

where the asterisks indicate complex conjugates. Integrating this equation over R $-$ δR, and supposing that the axial phase-change coefficient is the same for both fields, we find that the only contribution to the surface integral derived from the divergence comes from the normal component of the first term in it integrated over S, the part of the perfectly conducting surface of the perturbed structure which bounds δR. Thus

$$j(\omega_p - \omega)\int_{\mathrm{R}-\delta\mathrm{R}} d\mathrm{R}(\mu_0\boldsymbol{H}\cdot\boldsymbol{H}_p^* + \epsilon_0\boldsymbol{E}\cdot\boldsymbol{E}_p^*) + \int_{\mathrm{S}} d\mathrm{S}\boldsymbol{n}\cdot\boldsymbol{H}_p^* \times \boldsymbol{E} = 0 \quad (30)$$

where $\boldsymbol{n}$ is a unit vector, normal to S and directed into δR. It is now possible to approximate in eqn. (30) by replacing $\boldsymbol{E}_p$ and $\boldsymbol{H}_p$ by $\boldsymbol{E}$ and $\boldsymbol{H}$, respectively, and taking the volume integral over R instead of R $-$ δR. Since the tangential component of $\boldsymbol{E}$ vanishes on the surface of the unperturbed structure, the surface integral may be extended over the entire surface of δR; then it can be converted into a volume integral taken over δR. Finally, using eqns. (25) and (26), we obtain simple result

$$\frac{\omega_p - \omega}{\omega} = \frac{\int_{\delta\mathrm{R}} d\mathrm{R}(\mu_0|\boldsymbol{H}|^2 - \epsilon_0|\boldsymbol{E}|^2)}{\int_{\mathrm{R}} d\mathrm{R}(\mu_0|\boldsymbol{H}|^2 + \epsilon_0|\boldsymbol{E}|^2)} \quad . \quad . \quad (31)$$

Alternative proofs of the corresponding theorem for cavity resonators have been given by Slater[10] and Casimir.[11]

The perturbation of the dispersion curve of the rectangular structure, which is discussed in Section 2, is easily understood in the light of this theorem except in the case of the single-T structure for small values of the axial phase-change coefficient. The peculiar behaviour here comes about because the fast and slow waves are coupled together by the perturbation.

621.372.2

Paper No. 2295 R
Mar. 1957

THE COUPLING IMPEDANCE OF TAPE STRUCTURES

By P. N. BUTCHER, Ph.D.

(*The paper was first received* 20*th June, and in revised form* 17*th October*, 1956.)

SUMMARY

A periodic array of parallel straight tapes is considered. This can propagate a variety of TEM waves, each one corresponding to a different mode of excitation of the tapes. The case in which there is simply a fixed phase-change from one tape to the next is treated in detail. The exact field distribution is determined by using the theory of analytic functions of a complex variable. It is applied to the calculation of the coupling impedances and dispersion curves of tape helices, ladder lines, interdigital lines and meander lines. The calculations show that, with certain provisos, the product of the coupling impedance and the group velocity is the same for all these structures when they have the same tape length, tape width and gap width. The tape length is to be interpreted as the length of one turn in helices and the length of half a turn in meander lines. The provisos are as follows: the phase-change coefficient must be the same in each case; the coupling impedance must be evaluated at the array in helices, averaged over the array in ladder lines and evaluated at the centre lines of the array in interdigital and meander lines; finally, the ladder lines considered must be slightly perturbed easitron structures. Averaging over the array in helices does not alter the coupling impedance. Averaging over the array in interdigital and meander lines reduces the coupling impedance by a factor between 0·5 and 1. In a strongly perturbed easitron structure, the product of the average coupling impedance at the array and the group velocity differs from that in a slightly perturbed easitron structure by a factor for which an approximate formula is given. The use of tape structures in travelling-wave tubes for the millimetre wavebands is discussed briefly.

LIST OF PRINCIPAL SYMBOLS

A = Tape length.
c = Velocity of light in free space.
D = Period of the tape array.
E = Electric vector.
E_x, E_y, E_z = Cartesian components of the electric vector.
$E(\theta)$ = The z component of the space harmonic of the electric field which has phase-change θ per period of the array, evaluated on the plane $z = 0$.
$F(\theta)$ = Field distribution factor (f.d.f.).
$\mathrm{f}(\eta) = \dfrac{D}{\pi}\sqrt{\left(\sin^2\dfrac{\pi s}{2D} - \sin^2\dfrac{\pi\eta}{D}\right)}$.
G = Coefficient in the electric potential function for the interdigital line.
H = Magnetic vector.
H_x, H_y, H_z = Cartesian components of the magnetic vector.
I = Total current flowing along the tape to the right of the origin across the plane $x = 0$, in the direction of propagation, when the array propagates a TEM travelling wave.
$\mathrm{J}_0(\alpha)$ = Bessel function of order zero.
$K(\theta)$ = Coupling impedance.
$\mathrm{K}(\kappa)$ = Complete elliptic integral of the first kind with modulus κ.
k = Phase-change coefficient in free space.
$\mathrm{L}(\alpha) = [1 - 2/\pi \cos\alpha \mathrm{K}(\sin\alpha)][\mathrm{J}_0(\alpha) - \cos\alpha]^{-1}$.
n = Integer.
P = Mean power flowing along the guide.
r = Ridge factor.
s = Gap width.
t = Real variable of integration.
$U, U', U(\theta)$ = Electric potential function on the plane $x = 0$, when the array propagates a TEM travelling wave with phase change θ per period.
U_1, U_2 = Real and imaginary parts of U.
$U_0, U_0(\theta)$ = Value of $U(\theta)$ on the tape to the right of the origin.
u_n = Coefficients in the expansion of U as a series of space harmonics.
$\Delta U(s/D, \theta)$ = Difference of the values of $U(\theta)$ on the tapes on either side of the origin.
$V, V', V(\theta)$ = Magnetic potential function on the plane $x = 0$, when the array propagates a TEM travelling wave with phase-change θ per period.
V_1, V_2 = Real and imaginary parts of V.
v_n = Coefficients in the expansion of V as a series of space harmonics.
v_g = Group velocity.
w = Tape width.
W = Total energy stored per unit length of the guide.
W_1, W_2 = $U_1 + jV_1$ and $U_2 + jV_2$, respectively.
x, y, z = Cartesian co-ordinates.
$Z(\theta)$ = U_0/I, the characteristic impedance of a TEM travelling wave.
Z_0 = Intrinsic impedance of free space.
$\alpha = \pi s/(2D)$.
β_n = Phase-change coefficient of the nth space harmonic.
ϵ_0 = Permittivity of free space.
$\zeta = z + jy$.
η = Complex variable of integration.
θ = Phase-change per period of array.
$\mu = j\exp(j\theta/2)$.
μ_0 = Permeability of free space.
ρ = Distance from the edge of a tape.
τ = Real variable of integration.
ϕ, ϕ' = Electric potential functions.
ψ, ψ' = Magnetic potential functions.

(1) INTRODUCTION

Slow-wave structures which incorporate a periodic array of parallel metal tapes are finding many applications in travelling-wave tubes, particularly in the millimetre wavebands. Monofilar helices are useful in both forward-wave and backward-wave tubes.[1,2] Bifilar helices are particularly appropriate for backward-wave operation.[3] Tape ladder lines have been used in backward-wave oscillators for the millimetre wavebands;[4,5] they can be made easily either by photo-etching a thin plate or by winding tape on to a rod with a suitable channel cut out of it. Tape interdigital and meander lines for the millimetre wavebands could also be made easily by photo-etching, although no one has done so at the time of writing.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Butcher is now at the Stanford Electronics Research Laboratories, Stanford University, California, on leave from the Radar Research Establishment, Malvern.

The design of a travelling-wave tube rests largely on two basic features of the slow-wave structure: the dispersion curve and the coupling impedance. The principal object of the paper is to give a unified treatment of the dispersion curves and coupling impedances of tape structures. The discussion of the coupling impedance is made much easier by focusing attention on the product of the coupling impedance and the group velocity rather than on the coupling impedance itself. More precisely, the 'field distribution factor' (f.d.f.) is defined as

$$F(\theta) = \frac{K(\theta)}{Z_0}\frac{|v_g|}{c} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where θ is the phase change, per period of the tape array, of the space harmonic considered, $K(\theta)$ is the coupling impedance, v_g is the group velocity, and Z_0 and c are the free-space values of the intrinsic impedance and the velocity of light, respectively. The coupling impedance is given by the well-known formula[1]

$$K(\theta) = \frac{|E(\theta)|^2}{2(\theta/D)^2|P|} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where $|E(\theta)|$ is the amplitude of the axial electric field of the space harmonic considered, D is the period of the tape array and P is the mean power flowing through the structure. Substituting $K(\theta)$ into eqn. (1) and expressing P as the product of the group velocity and the mean stored energy per unit length of the guide, W, it is found that

$$F(\theta) = \frac{\epsilon_0|E(\theta)|^2}{2(\theta/D)^2W} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where ϵ_0 is the permittivity of free space. The main problem in calculating the coupling impedance is the evaluation of $F(\theta)$. The group velocity can be obtained easily from the dispersion curve.

The theory is based on the exact solution to the problem of TEM-wave propagation along a periodic array of parallel straight tapes. This is derived by using the theory of analytic functions of a complex variable in a way which is similar to the familiar technique for solving two-dimensional potential problems. However, in the present case the electromagnetic potential functions are generally complex and cannot be regarded as the real parts of suitable analytic functions. It is shown how this difficulty may be overcome. The M.K.S. system of units is used throughout the paper.

(2) TEM WAVES ON A PERIODIC ARRAY OF PARALLEL STRAIGHT TAPES

Consider the periodic array of parallel straight tapes shown in Fig. 1 having gap with s, tape width w and period D (so that $s + w = D$). It is convenient to use a right-handed system of rectangular Cartesian co-ordinate axes $0xyz$ located and oriented as shown in the Figure. The array can propagate a variety of TEM waves, each one corresponding to a different mode of excitation of the tapes. The relevant case is when there is simply a phase-change θ from one tape to the next. The electric and magnetic vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ can be derived from the electric and magnetic potential functions ϕ and ψ through the equations

$$\left.\begin{aligned} E_y &= -\frac{\partial\phi}{\partial y}, \quad Z_0H_y = -\frac{\partial\psi}{\partial y} \\ E_z &= -\frac{\partial\phi}{\partial z}, \quad Z_0H_z = -\frac{\partial\psi}{\partial z} \end{aligned}\right\} \quad . \quad . \quad . \quad (4)$$

For a wave travelling in the negative x direction,

$$\left.\begin{aligned} \phi &= U(y, z)\exp(jkx) \\ \psi &= V(y, z)\exp(jkx) \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (5)$$

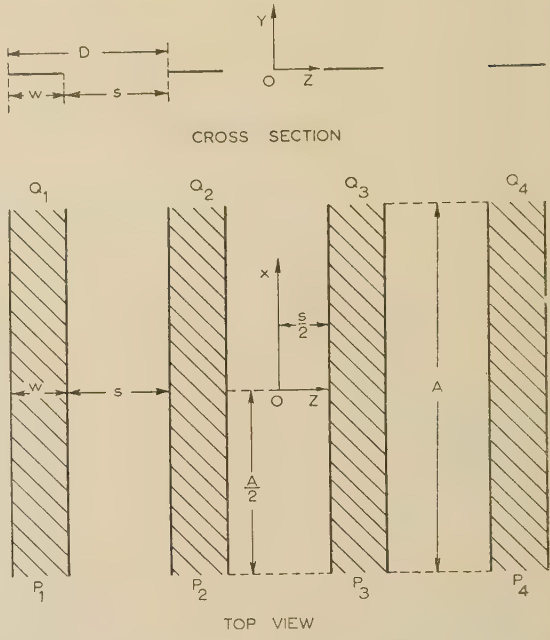


Fig. 1.—A periodic array of parallel straight tapes.

where U and V are the electric and magnetic potential functions on the plane $x = 0$, and k is the phase-change coefficient in free space at the frequency considered. Since $Z_0\boldsymbol{H}$ can be obtained from $\boldsymbol{E}$ by a right-handed rotation of 90° about the negative x direction, U and V are related by the equations

$$\left.\begin{aligned} \frac{\partial U}{\partial z} &= \frac{\partial V}{\partial y} \\ \frac{\partial U}{\partial y} &= -\frac{\partial V}{\partial z} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Moreover, they must satisfy the following boundary conditions:

$$\left.\begin{aligned} U(y, z + D) &= U(y, z)\exp(-j\theta) \\ V(y, z + D) &= V(y, z)\exp(-j\theta) \end{aligned}\right\} \quad . \quad . \quad (7a)$$

$$\left.\begin{aligned} &U, V \to 0, \ y \to \infty \\ &\nabla U, \nabla V \sim \rho^{-1/2} \text{ near the tape edges} \end{aligned}\right\} \quad . \quad . \quad (7b)$$

$$\frac{\partial U}{\partial z}, \frac{\partial V}{\partial y} = 0 \text{ on the tapes} \quad . \quad . \quad . \quad . \quad (7c)$$

$$\nabla U, \nabla V \text{ must be continuous except at the tapes} \quad . \quad (7d)$$

In the second part of eqn. (7b), ρ denotes the distance from the edge of the tape considered. The first boundary condition ensures that there is a phase change θ per period of the array; the second that the field falls to zero at infinity and has the appropriate singularities at the edges of the tapes; the third that the tangential electric field and normal magnetic field vanish on the tapes; and the fourth boundary condition ensures that the field vectors are continuous, except at the tapes.

It is easy to see that U is an even function of y, and V an odd function of y. Thus, from the symmetry of the array, it is clear that U is either even or odd, and from eqns. (6) it follows that V is correspondingly either odd or even. But U cannot be odd because then $\partial U/\partial z$ (and E_z) would be zero over the entire plane $y = 0$; hence the gaps between the tapes could be closed up without disturbing the field distribution. This would leave a TEM wave guided along an infinite perfectly conducting sheet

which obviously does not guide waves at all. Because of this symmetry $\partial U/\partial z$ and $\partial V/\partial y$ are necessarily continuous through the gaps between the tapes, while $\partial U/\partial y$ and $\partial V/\partial z$ will be continuous through the gaps provided that they vanish there. Thus the gaps are 'magnetic walls' on which the tangential magnetic field vanishes. It is convenient, therefore, to consider only the region above the array ($y > 0$) and to impose this extra boundary condition

$$\frac{\partial U}{\partial y}, \frac{\partial V}{\partial z} = 0 \text{ in the gaps} \quad . \quad . \quad (7e)$$

The solution to this boundary value problem is easily obtained once the right point of view is adopted. It will be observed that, apart from the fact that U and V are generally complex, eqns. (6) are identical with the familiar Cauchy–Riemann equations. Hence, separating U and V into real and imaginary parts by writing

$$\left.\begin{aligned} U &= U_1 + jU_2 \\ V &= V_1 + jV_2 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (8)$$

it follows immediately that U_1 and V_1 satisfy the Cauchy–Riemann equations, and so do U_2 and V_2. Thus

$$\left.\begin{aligned} W_1 &= U_1 + jV_1 \\ W_2 &= U_2 + jV_2 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (9)$$

are analytic functions of the complex variable $\zeta = z + jy$. The boundary conditions for W_1 and W_2 can be obtained from the boundary conditions (7) for U and V. They are:

$$\left.\begin{aligned} W_1(\zeta + D) &= \cos\theta W_1(\zeta) + \sin\theta W_2(\zeta) \\ W_2(\zeta + D) &= \cos\theta W_2(\zeta) - \sin\theta W_1(\zeta) \end{aligned}\right\} \quad . \quad (10a)$$

$$\left.\begin{aligned} &W_1, W_2 \to 0, y \to \infty \\ &\frac{dW_1}{d\zeta}, \frac{dW_2}{d\zeta} \sim \rho^{-1/2} \text{ near the tapes edges} \end{aligned}\right\} \quad . \quad (10b)$$

$$\left.\begin{aligned} &\frac{dW_1}{d\zeta}, \frac{dW_2}{d\zeta} \text{ must be imaginary on the tapes} \\ &\frac{dW_1}{d\zeta}, \frac{dW_2}{d\zeta} \text{ must be real in the gaps} \end{aligned}\right\} \quad . \quad (10c)$$

A pair of analytic functions which satisfy these boundary conditions can be determined by trial, using as a starting-point the well-known solution for the case $\theta = \pi$ (see Reference 6). They are

$$W_1(\zeta) = \int_{\infty}^{\zeta} d\eta \frac{\cos(\pi - \theta)\frac{\eta}{D}}{f(\eta)} \quad . \quad . \quad . \quad (11a)$$

$$W_2(\zeta) = \int_{\infty}^{\zeta} d\eta \frac{\sin(\pi - \theta)\frac{\eta}{D}}{f(\eta)} \quad . \quad . \quad . \quad (11b)$$

when θ is positive and less than 2π. The contour of integration is to start at $y = \infty$ and remain above the array. The function f(η) is

$$f(\eta) = \frac{D}{\pi}\sqrt{\left(\sin^2\frac{\pi s}{2D} - \sin^2\frac{\pi\eta}{D}\right)} \quad . \quad . \quad . \quad (12)$$

It is defined to be real and positive in the gap containing the origin of co-ordinates and is continued analytically from there. The integrals do not converge when θ is negative or greater than 2π. However, the range 0 to 2π includes all the physically distinct cases.

There are several things to notice about f(η). It is real in the gaps and imaginary on the tapes. More particularly, f(η) has phases $-\pi/2$ and $+\pi/2$ on the tapes to the right and left of the origin, respectively. (This can be shown by following the phase of f(η) round two contours in the upper half-plane which start at the origin and finish on one or the other of these tapes.) Finally, it is clear from eqn. (12) that $f(\eta + D) = \pm f(\eta)$; now, if η lies on the tape to the left of the origin, then $\eta + D$ lies on the tape to the right of the origin and, from the previous remarks, the minus sign is appropriate. Thus, for any value of η,

$$f(\eta + D) = - f(\eta) \quad . \quad . \quad . \quad . \quad . \quad (13)$$

With these facts in mind, it is easy to verify that W_1 and W_2 satisfy the boundary conditions (10).

One can now derive the field vectors. We need only consider E_z on the line $x = 0$ in the gap containing the origin of co-ordinates and H_z on the line $x = 0$ on the tape to the right of the origin. In the gap, $dW_1/d\zeta$ and $dW_2/d\zeta$ are real and equal to $\partial U_1/\partial z$ and $\partial U_2/\partial z$, respectively. Hence, using eqns. (4), (5), (8), (11) and (12),

$$\left.\begin{aligned} E_z(0, 0, z) &= -\left(\frac{\partial U_1}{\partial z} + j\frac{\partial U_2}{\partial z}\right) \\ &= -\left(\frac{dW_1}{d\zeta} + j\frac{dW_2}{d\zeta}\right)_{\zeta=z} \\ &= -\frac{\exp j(\pi - \theta)z/D}{\frac{D}{\pi}\sqrt{\left(\sin^2\frac{\pi s}{2D} - \sin^2\frac{\pi z}{D}\right)}} \end{aligned}\right\} \quad . \quad (14)$$

where the positive square root must be taken. On the tape to the right of the origin, $dW_1/d\zeta$ and $dW_2/d\zeta$ are imaginary and equal to $j\partial V_1/\partial z$ and $j\partial V_2/\partial z$, respectively. Hence, using eqns. (4), (5), (8) and (11),

$$\left.\begin{aligned} H_z(0, 0, z) &= -\frac{1}{Z_0}\left(\frac{\partial V_1}{\partial z} + j\frac{\partial V_2}{\partial z}\right) \\ &= -\frac{1}{Z_0 j}\left(\frac{dW_1}{d\zeta} + j\frac{dW_2}{d\zeta}\right)_{\zeta=z} \\ &= \frac{j}{Z_0}\frac{\exp j(\pi - \theta)z/D}{f(z)} \end{aligned}\right\} \quad . \quad (15)$$

Now, f(z) has phase $-\pi/2$ on this tape so that, from eqn. (12),

$$\left.\begin{aligned} f(z) &= -\frac{jD}{\pi}\sqrt{\left(\sin^2\frac{\pi z}{D} - \sin^2\frac{\pi s}{2D}\right)} \\ &= -j\frac{D}{\pi}\sqrt{\left(\sin^2\frac{\pi w}{2D} - \cos^2\frac{\pi z}{D}\right)} \end{aligned}\right\} \quad . \quad . \quad (16)$$

since $s + w = D$. The positive square root must be taken. Substitution in eqn. (15) gives, finally,

$$H_z(0, 0, z) = \frac{-\exp j(\pi - \theta)z/D}{\frac{Z_0 D}{\pi}\sqrt{\left(\sin^2\frac{\pi w}{2D} - \cos^2\frac{\pi z}{D}\right)}} \quad . \quad . \quad (17)$$

The similarity of the formulae for $E_z(0, 0, z)$ and $H_z(0, 0, z)$ is by no means accidental. It is a consequence of Babinet's principle. The application of this principle to tape structures leads to several useful and interesting results. This topic has been relegated to an Appendix, Section 11.1, so as not to interrupt the present discussion.

(3) THE CHARACTERISTIC IMPEDANCE AND FIELD DISTRIBUTION FACTOR OF A TEM WAVE

Following Fletcher,[7] we define the characteristic impedance, $Z(\theta)$, of an array propagating a TEM travelling wave with phase change θ per period as the ratio, for any tape, of the electric potential to the total current flowing in the direction of propagation. Consider the TEM travelling wave which was discussed in the previous Section. Attention will be focused on the plane $x = 0$. If U_0 is the electric potential on the tape to the right of the origin, then the electric potential on the tape to the left of the origin is $U_0 \exp j\theta$ and the potential difference is $U_0(1 - \exp j\theta)$ Hence U_0 is given by

$$U_0(1 - \exp j\theta) = \Delta U(s/D, \theta)$$

i.e.

$$U_0 = \frac{j \exp(-j\theta/2)}{2 \sin \theta/2} \Delta U(s/D, \theta) \quad . \quad . \quad . \quad (18)$$

where

$$\Delta U(s/D, \theta) = \int_{s/2}^{-s/2} dz\, E_z(0, 0, z) \quad . \quad . \quad . \quad (19a)$$

Using eqn. (14), making the change of variable $t = -\pi z/D$ and noticing that the imaginary part of the integrand is an odd function of t, it is found that

$$\Delta U(s/D, \theta) = \int_{-\pi s/(2D)}^{\pi s/(2D)} dt \frac{\cos(1 - \theta/\pi)t}{\sqrt{\left(\sin^2 \frac{\pi s}{2D} - \sin^2 t\right)}} \quad . \quad (19b)$$

This integral is evaluated approximately in Section 11.2.

The total current I flowing along the tape to the right of the origin across the plane $x = 0$, in the negative x direction, is equal to twice the current flowing in the upper surface of the tape. Thus,

$$I = 2 \int_{s/2+w}^{s/2} dz\, H_z(0, 0, z) \quad . \quad . \quad . \quad . \quad (20a)$$

Using eqn. (17), making the change of variable $t = \pi/2 - \pi z/D$ and noticing that part of the integrand is an odd function of t, we obtain

$$I = \frac{2}{Z_0} j \exp(-j\theta/2) \Delta U(w/D, \theta) \quad . \quad . \quad . \quad (20b)$$

where $U(w/D, \theta)$ is given by eqn. (19b) with s replaced by w. It follows immediately from eqns. (18) and (20b) that the characteristic impedance is

$$Z(\theta) \equiv \frac{U_0}{I} = \frac{Z_0}{4 \sin(\theta/2)} \frac{\Delta U(s/D, \theta)}{\Delta U(w/D, \theta)} \quad . \quad . \quad (21a)$$

Notice that when $s = w$ the last factor on the right-hand side of this equation reduces to unity, yielding the extremely simple formula

$$Z(\theta) = \frac{Z_0}{4 \sin(\theta/2)}, \quad s = w \quad . \quad . \quad . \quad . \quad (21b)$$

It only remains to calculate the f.d.f. (which is defined in Section 1) of a section of the array, with length A in the x direction, when it supports a TEM travelling wave. This is very closely linked to the f.d.f. of structures which incorporate such a section of the array. The z-component of the electric field of the space harmonic with phase change θ, per period of the array, is $E(\theta)$, $\exp(-j\theta z/D)$, where

$$E(\theta) = \frac{1}{D} \int_{-D/2}^{D/2} dz\, E_z(x, y, z) \exp(j\theta z/D) \quad . \quad . \quad (22)$$

Setting $E_z(0, 0, z)$ equal to zero on the tapes and equal to the right-hand side of eqn. (14) in the gap, noticing that the imaginary part of the integrand is an odd function of z, and making the change of variable $t = \sin \pi z/D$, it is found (after a trivial integration) that

$$E(\theta) = -\pi/D \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

when $x = 0$, $y = 0$ and θ lies between 0 and 2π. Turning to the calculation of W, we observe that the mean power flowing in a period of the array *in the negative x-direction* is $\frac{1}{2}|U_0|^2/Z(\theta)$. Moreover, the velocity of group propagation *in the negative-x direction* is simply the velocity of light in free space. It follows that the mean energy stored in a period of the array per unit length *in the x-direction* is $\frac{1}{2c}|U_0|^2/Z(\theta)$. Hence the mean energy stored in length A in the x-direction per unit length *in the z-direction* is

$$W = \frac{1}{D} \frac{A}{2c} \frac{|U_0|^2}{Z(\theta)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (24a)$$

$$= \frac{\epsilon_0 A}{2D} \frac{\Delta U(s/D, \theta) \Delta U(w/D, \theta)}{\sin \theta/2} \quad . \quad . \quad (24b)$$

from eqns. (18) and (21a). Finally, substitution of eqns. (23) and (24b) into eqn. (3) gives the f.d.f. at the array

$$F(\theta) = \frac{D}{A} \frac{\pi^2 \sin \theta/2}{\theta^2 \Delta U(s/D, \theta) \Delta U(w/D, \theta)} \quad . \quad . \quad (25)$$

Inspection of this equation shows that $F(\theta)$ is unaltered when the slots and tapes are interchanged. This is a consequence of Babinet's principle (see Section 11.1). The f.d.f. is infinite when $\theta = 0$ and zero when $\theta = 2\pi$.

In Fig. 2, $F(\theta)$ is plotted against s/D for $\theta = \pi/4, \pi/3, \pi/2, \pi$

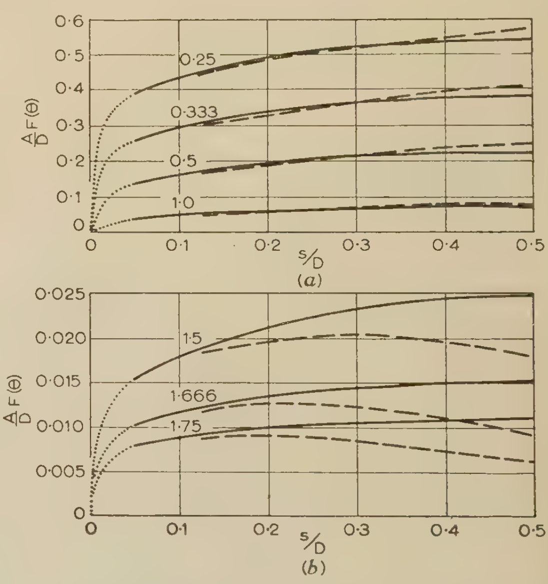
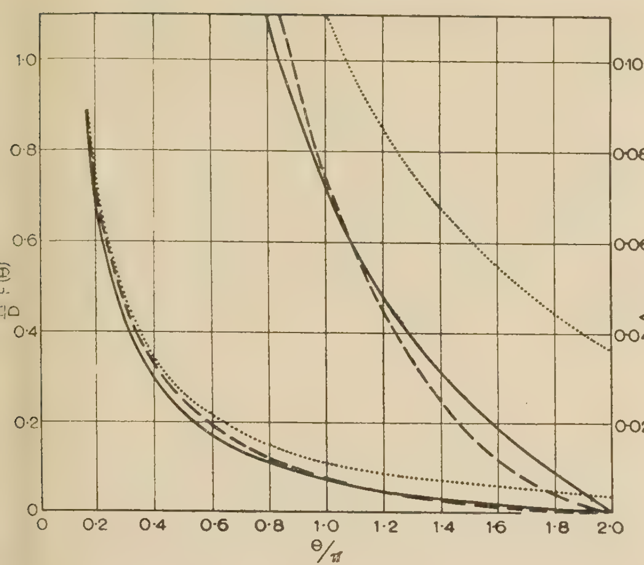


Fig. 2.—The field distribution factor of a TEM travelling wave as a function of the gap width.

—— Exact value (extrapolated to the origin by eye).
---- Constant-field approximation.
θ/π has the values indicated on the curves.

$3\pi/2$, $5\pi/3$, and $7\pi/4$. In every case there is a very broad maximum at $s/D = 0{\cdot}5$. In Fig. 3, $F(\theta)$ is plotted against θ for $s/D = 0{\cdot}5$. The Figures also show the corresponding values of $F(\theta)$ obtained on the basis of the 'constant-field approximation', in which $E_z(0, 0, z)$ is taken to be constant within each gap (see



ig. 3.—The field distribution factor of a TEM travelling wave as a function of the phase change per period of the array.

—— Exact value of $(A/D)F(\theta)$.
---- Constant-field approximation to $(A/D)F(\theta)$.
····· Exact value of $(A/D)\frac{1}{2}F(\theta/2)$.

The gap width is equal to the tape width in every case. Part of each curve is drawn ain on the larger scale indicated on the right-hand side of the graph.

ection 11.3). This approximation entails a considerable error hen θ/π is greater than $1\cdot3$ unless the gaps are very narrow. or small θ, however, the error is negligible. The reason for is is as follows. The space harmonic with phase change θ per eriod of the array falls off exponentially in the y-direction with n attenuation coefficient θ/D. Thus it extends a long way from e array when θ is small. Most of the field energy is then ored in this space harmonic and $F(\theta)$ approaches the value $/2A\theta$ which is obtained by ignoring all the other space harmonics (see Section 11.3). Thus $F(\theta)$ is independent of the field stribution in the gaps when θ is small. It is surprising, perhaps, at the constant-field approximation still gives excellent results hen θ is in the neighbourhood of π.

(4) HELICES

A section of the array, with length A in the x-direction, can be garded as a development of either a monofilar or a multifilar lix in which the length of one turn is A and the period is D. eferring to Fig. 1, it is seen that when the array is wrapped und a right circular cylinder so as to form a monofilar helix, is joined to Q_2, P_2 to Q_3, P_3 to Q_4, and so on down the array. n the other hand, when we form a bifilar helix, P_1 is joined to $_3$, P_2 to Q_4, and so on. The well-known dispersion equations these two cases are obtained by requiring that the electric otentials and total tape currents should match when these ins are made. Consider a TEM wave travelling in the negative positive x-direction with phase change θ per period of the ray. The total phase advance between P_1 and Q_2 is $\pm kA - \theta$, d between P_1 and Q_3 it is $\pm kA - 2\theta$. Thus the dispersion rves are given by the equations

$$\pm kA - \theta = 2n\pi \quad . \quad . \quad . \quad . \quad . \quad (26)$$

r the monofilar helix, and

$$\pm kA - 2\theta = 2n\pi \quad . \quad . \quad . \quad . \quad . \quad (27)$$

r the bifilar helix. Here n is an arbitrary integer. The branches of the dispersion curves for which θ lies between 0 and 2π are plotted in Figs. 4(*a*) and 4(*b*). In the bifilar case, the full lines correspond to the push-pull mode and the dashed lines correspond to the push-push mode.[3]

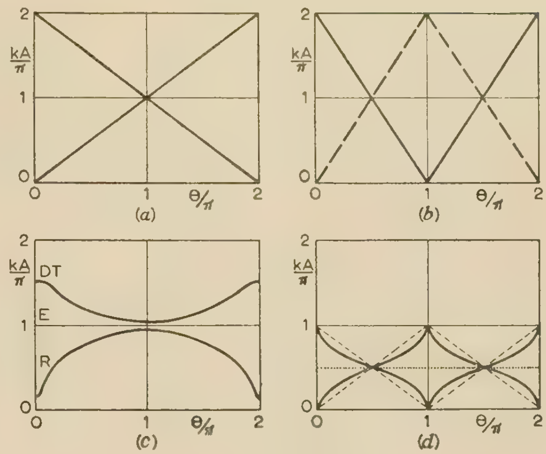


Fig. 4.—Sketch of the dispersion curves.

(*a*) Monofilar helix.
(*b*) Bifilar helix; the solid lines and dashed lines correspond to the push-pull and the push-push modes, respectively.
(*c*) Ladder lines; double-T structure (DT), easitron structure (E), ridge structure (R).
(*d*) Interdigital and meander lines; the solid curves, dashed lines and dotted lines are (i) the dispersion curves of an interdigital line in which $s = w$, $s \ll w$ and $s \gg w$, respectively; (ii) the dispersion curves of a meander line in which $s = w$, $s \gg w$ and $s \ll w$, respectively.

It should be obvious that the f.d.f. at the array in either case is simply $F(\theta)$ as calculated in the previous Section and plotted in Figs. 2 and 3. Notice, therefore, that the f.d.f. is determined entirely by the dimensions of the structure and the value of θ. It does not matter whether one is concerned with the forward wave in the monofilar helix, the backward wave in the monofilar helix, the push-push mode in the bifilar helix or the push-pull mode in the bifilar helix. The calculation of the coupling impedance is completed by the evaluation of $c/|v_g| = |\partial\theta/\partial kA|(A/D)$ from eqns. (26) and (27). It is found immediately that $c/|v_g|$ has the value A/D in the monofilar helix and $A/(2D)$ in the bifilar helix.

(5) LADDER LINES

The simplest ladder line is the easitron structure* in which a length A of the array is short-circuited at each end by an infinite perfectly conducting plate. TEM standing waves can be set up at the frequencies for which the free-space wavelength is twice the tape length. Thus $kA = \pi$, for all θ, and the dispersion curve is the horizontal straight line labelled E in Fig. 4(*c*). The 'incident' and 'reflected' TEM travelling waves have equal amplitudes which may be chosen so that the axial electric field at the centre line of the array is the same as that of the travelling wave considered in Sections 2 and 3. Then, at the centre line of the array, $E(\theta) = -\pi/D$ as in eqn. (23). The stored energy per unit length has, however, half the value given by eqn. (24) because of the sinusoidal variation of the field vectors in the x-direction. Thus the f.d.f. at the centre line of the array is equal to twice the value given by eqn. (25). The *average* value of the f.d.f. at the array is precisely equal to $F(\theta)$ as given by eqn. (25) because of the sinusoidal variation of $E(\theta)$ in the x-direction.

The group velocity of the slow wave in the easitron structure is zero. The slow wave can be made to propagate energy by inserting ridges above and below the array either at the centre of the guide (ridge structure) or at the sides of the guide

* This terminology is adopted because a ladder line of this type was used by Walker at the Bell Telephone Laboratories in a tube called an 'easitron' (unpublished).

(T-structure). The dispersion curves of a ridge structure and a double-T structure are sketched in Fig. 4(*c*) and labelled R and DT, respectively. The group velocity is now a rapidly varying function of θ, falling to zero at $\theta = \pi$. The dispersion curves of ridge and T-structures are treated in detail in the previous paper.[8] The f.d.f. in ridge and T-structures is, of course, different from that in the corresponding easitron structure. The appropriate correction factors are roughly estimated in Section 11.4.

(6) INTERDIGITAL AND MEANDER LINES

In an interdigital line, P_1 is short-circuited to P_3 and Q_2 is short-circuited to Q_4, while P_2, P_4, Q_1 and Q_3 are open-circuited, and so on down the array. The period of the whole structure is $2D$, but it is also unchanged when it is moved along the z-axis through half a period and then reflected in the plane $x = 0$. It is possible, therefore, to consider a mode for which the electric field at $(-x, y, z + D)$ differs from that at (x, y, z) only by a constant factor $\exp(-j\theta)$. Clearly θ is the phase change of the electric field on the plane $x = 0$ per period of the tap array. The effect of applying the above operation twice is simply to move the structure along the z-axis through a complete period $2D$. Hence the phase change per period of the interdigital line is 2θ.

The field about the array is a TEM standing wave. However, because of the screw symmetry it is more complicated than that in the easitron structure. If θ lies between 0 and π, the appropriate electric potential function is

$$\phi = U(\theta)\,[\exp(jkx) + \exp(-jkx)] + GU(\theta + \pi)\,[\exp(jkx) - \exp(-jkx)] \quad . \quad (28)$$

where $U(\theta)$ is defined by eqns. (8), (9) and (11), $U(\theta + \pi)$ is defined by the same equations with θ replaced by $\theta + \pi$, and G is some constant.

The dispersion equation and the value of G are determined by the boundary conditions at P_3 and Q_3. The points P_3 and P_1 are short-circuited so that the electric potential difference between them is zero. But the electric potential at P_1 is just $\exp(2j\theta)$ times that at P_3. Hence the electric potential at P_3 must itself vanish, i.e. from eqn. (28),

$$U_0(\theta)\,[\exp(-jkA/2) + \exp(jkA/2)] + GU_0(\theta + \pi)[\exp(-jkA/2) - \exp(jkA/2)] = 0 \quad . \quad (29)$$

where $U_0(\theta)$ and $U_0(\theta + \pi)$ are respectively the values of $U(\theta)$ and $U(\theta + \pi)$ on the tape to the right of the origin of co-ordinates. The point Q_3 is open-circuited, i.e. the total current there is zero. Hence, from eqns. (28) and (21*a*),

$$\frac{U_0(\theta)}{Z(\theta)}\,[\exp(jkA/2) - \exp(-jkA/2)] + \frac{GU_0(\theta + \pi)}{Z(\theta + \pi)}[\exp(jkA/2) + \exp(-jkA/2)] = 0 \quad . \quad (30)$$

where it has been remembered that the current I, in eqn. (21*a*), flows in the direction of propagation. Eqns. (29) and (30) give immediately the dispersion equation

$$\cot^2\left(\frac{kA}{2}\right) = \frac{Z(\theta + \pi)}{Z(\theta)} \quad . \quad . \quad . \quad . \quad (31)$$

and an equation for the amplitude of G,

$$\frac{|U_0(\theta)|^2}{Z(\theta)} = \frac{|GU_0(\theta + \pi)|^2}{Z(\theta + \pi)} \quad . \quad . \quad . \quad . \quad (32)$$

The dispersion equation is of some interest because tape interdigital lines have not been treated before. Substituting for $Z(\theta)$ and $Z(\theta + \pi)$ from eqn. (21*a*) into eqn. (31) gives

$$\cot^2\frac{kA}{2} = \tan\frac{\theta}{2}\left[\frac{\Delta U(w/D, \theta)\Delta U(s/D, \theta + \pi)}{\Delta U(s/D, \theta)\Delta U(w/D, \theta + \pi)}\right] \quad . \quad (33a)$$

Inspection shows that $kA = \pi$ when $\theta = 0$, and $kA = 0$ when $\theta = \pi$. Moreover, it is found from eqn. (19*b*) that the right-hand side of eqn. (33*a*) reduces to unity when $\theta = \pi/2$ and so $kA = \pi/2$. It also follows from eqns. (33*a*) and (19*b*) that $kA - \pi/2$ is an odd function of $\theta - \pi/2$. In order to obtain the full dispersion curve it is generally necessary to evaluate $\Delta U(w/D, \theta)$, etc. (see Section 11.2). However, in the particular case $s = w$, these quantities cancel out, and eqn. (33*a*) reads simply

$$\cot^2\left(\frac{kA}{2}\right) = \tan\frac{\theta}{2}, \quad s = w \quad . \quad . \quad . \quad (33b)$$

In both the eqns. (33), θ is restricted to lie between 0 and π (as in all the preceding analysis in this Section). The corresponding branch of the dispersion curve is plotted in Fig. 5 for $s/D = 0{\cdot}25$,

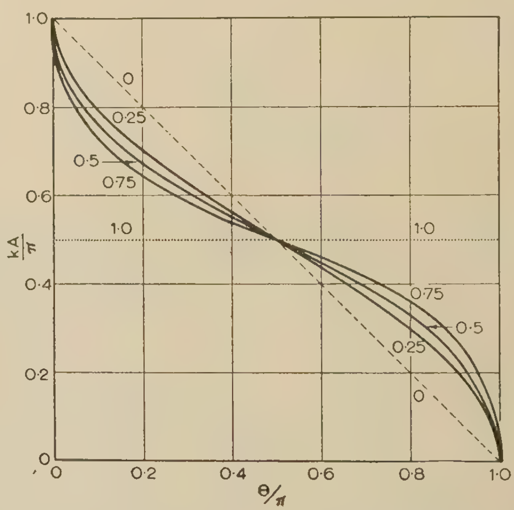


Fig. 5.—Dispersion curves for interdigital and meander lines.

The curves apply to both interdigital and meander lines provided that the number which labels each curve is interpreted as s/D in the first case and w/D in the second.

0·5 (i.e. $s = w$) and 0·75. Notice that it corresponds to a backward space harmonic. The complete dispersion curve can be obtained from this branch by displacing it by integral multiples of π along the θ-axis and then reflecting all these branches in the k-axis so as to obtain the branches corresponding to a wave travelling in the opposite direction. The complete dispersion curve is sketched in Fig. 4(*d*) for values of θ between 0 and 2π.

The group velocity can be obtained from the slope of the dispersion curve. It is useful to observe that $c/|v_g| = 2A/D$ at mid-band in the case $s = w$. This can be shown by differentiating eqn. (33*b*) with respect to θ and remembering that $\theta = \pi/2$ when $kA = \pi/2$. Notice that the group velocity tends to infinity when θ approaches any integral multiple of π. Inspection of the dispersion curve shows, however, that the points on it for which the group velocity has a magnitude greater than the velocity of light actually lie in 'forbidden regions',[9] such that the phase velocity of one of the space harmonics is also greater than the velocity of light. The exact dispersion curve of a completely open structure cannot pass through the forbidden regions because, if it did, the structure would radiate. We have not mentioned this complication before in the paper because it is not very important in practice (but see Reference 8). In the present case, however, it

nplies that the calculations break down near the edges of the ass band.

The tape interdigital line has a markedly non-linear dispersion urve because the thin fingers do not shield successive gaps from ne another very effectively. When the tapes are very wide, owever, successive gaps *are* shielded from one another, and each ap can be regarded as an individual transmission line propa-ating a TEM travelling wave. The succession of gaps can be ooked on as a folded transmission line in which the electric field ndergoes 180° phase change at each fold because of the geometry f that region. The various branches of this dispersion curve btained on this basis are given by

$$\pm kA - 2\theta = 2n\pi \quad . \quad . \quad . \quad . \quad . \quad (34)$$

here n is any integer. These equations define the dashed raight lines in Figs. 4(*d*) and 5. They also apply to interdigital nes in which successive gaps are shielded from one another ecause the fingers are thick in the y-direction.[10] Going to the pposite extreme, i.e. making the gaps very wide, isolates the ngers from one another. Each finger supports a TEM wave hich is short-circuited at one end and open-circuited at the her. Hence, whatever the value of θ, the free-space wavelength as the value $4A$. Thus we obtain the dotted straight lines in igs. 4(*d*) and 5. Inspection of Fig. 5 shows that the dispersion urves for these extreme cases are approached only very slowly s/D tends to 0 or 1, as the case may be.

We turn now to the field distribution factor. The electric otential function given in eqn. (28) is a superposition of four avelling waves. It can be shown that the corresponding electro-agnetic fields store energy independently of one another. oreover, from eqns. (24*a*) and (32), we see that the energy ored by each travelling wave per unit length in the z-direction equal to W as given by eqn. (24*b*). Hence the total energy r unit length in the z-direction is $4W$. However, the electric otential function is equal to $2U(\theta)$ on the centre line of the ray, and, on this line, $E(\theta)$ has twice the value given by eqn. (23). follows from eqn. (3) that the f.d.f. at the centre line of the ray is precisely $F(\theta)$ as given by eqn. (25).

From Fig. 5 it is clear that the f.d.f. calculated above applies the backward space harmonic for which θ lies between 0 and π. ne forward space harmonic, for which θ lies between 0 and π ig. 4(*d*)], has zero axial electric field on the centre line of the ray. The above analysis can be extended to values of θ lying tween π and 2π by replacing $\theta + \pi$ by $\theta - \pi$ everywhere. us we find that the f.d.f., on the centre line of the array, of the rward space harmonic for which θ lies between π and 2π is θ) as given by eqn. (25). It is now the backward space har-onic which has zero axial electric field on the centre line of e array.

For any value of θ between 0 and 2π, the f.d.f. with which we e concerned has its maximum value at the centre line of the ray. Averaging in the x-direction reduces the f.d.f. from its ak value by the factor

$$\frac{1}{A}\int_{-A/2}^{A/2} dx \cos^2 kx = \frac{1}{2}\left(1 + \frac{\sin kA}{kA}\right) \quad . \quad . \quad . \quad (35)$$

is varies from 1·0 at the bottom of the pass band to 0·5 at top. It is 0·82 at mid-band.

The meander line remains to be considered. In this case P_1 P_2, Q_2 and Q_3, and P_3 and P_4 are short-circuited, and so on the array. As in the interdigital line, the period of the ucture is $2D$ and it is unchanged when it is moved axially ough half a period and then reflected in the plane $x = 0$. If θ between 0 and π the appropriate electric potential function is ven by eqn. (28), but the value of G and the relation between θ and k will be different in this case. They are determined by the boundary conditions at P_3 and Q_3. Alternatively, one can make use of Babinet's principle (see Section 11.1). Thus it is found that the complete dispersion curve of a meander line with gap width s, tape width w and tape length A is the same as that of an interdigital with *gap* width w, *tape* width s and tape length A. The same is true of the f.d.f. at the centre line of the array. Moreover, since the right-hand side of eqn. (25) is unaltered when s and w are interchanged, the f.d.f. at the centre line of the array in a meander line is equal to that in an interdigital line with exactly the same dimensions as the meander line. In the meander line, however, the f.d.f. applies to the forward space harmonic when θ lies between 0 and π, and to the backward space harmonic when θ lies between π and 2π. It is the other way round in the interdigital line. Finally, averaging in the x-direction reduces the f.d.f. from its value on the centre line of the array by the factor given by eqn. (35), where k now has the value appropriate to the meander line for the value of θ considered.

(7) COMPARISON OF STRUCTURES

We have seen that, with certain provisos, the f.d.f. is the same in helices, easitron structures, interdigital lines and meander lines when they have the same tape length, tape width and gap width and the same value of θ is considered in each case. The tape length is to be interpreted as the length of one turn in helices and the length of half a turn in meander lines. The provisos are as follows: the f.d.f. must be evaluated at the array in helices, averaged over the array in the easitron structure, and evaluated at the centre line of the array in interdigital and meander lines. In the case of helices, averaging over the array does not alter the f.d.f., whereas, in the case of interdigital and meander lines, this averaging reduces the f.d.f. by a factor between 0·5 and 1. The perturbation of the easitron structure which must be introduced to make the slow wave propagate energy will modify the f.d.f., as discussed in Section 11.4.

We shall make a brief comparison of the relative merits of these various structures with regard to travelling-wave tube applications in the millimetre wavebands. It is convenient to label quantities referring specifically to any particular structure by a suffix. The suffices m, b, r, t, i and s will refer to the monofilar helix, the bifilar helix, the ridge ladder line, the T-ladder line, the interdigital line and the meander line, respectively. The discussion will be confined to structures with equal gap and tape widths. The f.d.f. at the centre line of the array is maximized in this case; there is nothing much to be gained by relaxing this restriction.

Consider forward-wave operation first. The monofilar helix is frequently used for this purpose. The phase and group velocities of the forward space harmonic with θ_m between 0 and 2π are equal, both having the value $c(D_m/A_m)$. The electrons can be coupled to this space harmonic at a beam voltage which is independent of the frequency. The f.d.f. at the helix is $F_m(\theta_m)$. The solid line in Fig. 3 is a plot of $F(\theta)$ as given by eqn. (25); it gives the f.d.f. at the helix in the monofilar case when $A_m = A$, $D_m = D$ and $kA = \theta_m = \theta$. In practice, the dimensions of the helix are usually chosen so that $kA < \pi/2$ because the f.d.f. is then relatively large.

In the bifilar helix the electrons can be coupled to the forward space harmonic of the push-push mode with θ_b between 0 and π. The beam voltage is independent of the frequency and is the same as that in the monofilar helix if $A_b = A_m$ and $D_b = D_m/2$, i.e. the pitch of the helices is the same in both cases. For the same frequency, $\theta_b = \theta_m/2$ and the f.d.f. at the bifilar helix is $F_b(\theta_b) = \frac{1}{2}F_m(\theta_m/2)$, from eqn. (25). The dotted line in Fig. 3 is a plot of $\frac{1}{2}F(\theta/2)$, where $F(\theta)$ is given by eqn. (25); it gives the f.d.f. at the helix in the bifilar case when $A_b = A$, $D_b = D/2$ and

$kA = 2\theta_b = \theta$. Clearly there is some advantage in using the bifilar helix instead of the monofilar helix, but it is not significant when $kA < \pi/2$.

The other structures which have been considered, ladder, interdigital and meander lines, have non-linear dispersion curves. Consequently they will be narrow-banded in forward-wave operation in comparison with helices. However, ladder and interdigital lines are more rugged than helices and have high thermal capacities. This is not true of meander lines, which are perhaps of academic interest only. Useful bandwidths can be achieved by synchronizing the electrons with a forward space harmonic which has equal phase and group velocities at the centre of the band. The synchronous beam voltage is then least sensitive to small changes of the frequency.

From the dispersion curves given in the previous paper[8] it appears that the phase and group velocities of the forward space harmonic in the ridge ladder line, with θ_r between 0 and π, are never equal; and from Figs. 4 and 5 of the present paper, the same is true of the forward space harmonic in the interdigital line with θ_i between π and 2π. These structures are therefore unsuitable for forward-wave operation unless one couples into a high-order space harmonic. Alternatively, in the interdigital line, it would be possible to couple into the forward space harmonic with θ_i between 0 and π. This has zero axial electric field on the centre line of the array, but the *average* f.d.f. at the array is finite, and the phase and group velocities are equal when $\theta_i = 0{\cdot}84\pi$.* This is not likely to be a very useful mode of operation.

In a suitably dimensioned T-ladder line, the phase and group velocities of the forward space harmonic with θ_t between π and 2π are equal for some value of θ_t. The average f.d.f. at the array is equal to that of a monofilar helix with the same period and tape length when $\theta_m = \theta_t$ (apart from the correction factor discussed in Section 11.4). The group velocity, beam voltage and frequency are all less than they are in the monofilar helix when $\theta_m = \theta_t$. It should be remarked, however, that the frequency is several times larger than that at which the helix would normally be used. Scaling down the helix until $kA_m < \pi/2$ at the frequency of operation of the T-ladder line would improve the performance of the helix considerably.

In the meander line, the phase and group velocities of the forward space harmonic are equal when $\theta_s = 0{\cdot}84\pi$ and $kA_s = 2{\cdot}2$.* The average f.d.f. at the array is about 68% of that in a monofilar helix, with the same dimensions, when $\theta_m = \theta_s$. The remarks made above about the group velocity, beam voltage and frequency in the T-ladder line apply to the meander line as well.

It remains to consider backward-wave operation. The monofilar helix has been used for this purpose.[2] The electrons are coupled to the backward space harmonic with θ_m between 0 and 2π. The rate of change of frequency with beam voltage (the tuning rate) is highest at low voltages, i.e. when θ_m is close to 2π. However, it will be seen from Fig. 3 that the f.d.f. tends to zero when θ_m tends to 2π. For this reason, the bifilar helix with $A_b = A_m$ and $D_b = D_m/2$ is superior to the monofilar helix.[3] The electrons can be coupled to the backward space harmonic of the push-pull mode with θ_b between 0 and π. The f.d.f. at any frequency is increased over that in the monofilar helix in the same way as was noted in the case of forward-wave operation. The improvement is much more significant in the case of backward-wave operation, particularly when θ_m approaches 2π (see Fig. 3; the solid line gives the f.d.f. at the monofilar helix when $A_m = A$, $D_m = D$ and $kA = 2\pi - \theta$; the dotted line gives the f.d.f. at the bifilar helix when $A_b = A$, $D_b = D/2$ and $kA = 2\pi - \theta$).

* This conclusion may not be valid because, as mentioned in Section 6, the calculations break down near the edges of the pass band.

In ridge ladder lines the electrons can be coupled to the reverse space harmonic with θ_r between π and 2π. The group velocity is a rapidly varying function of the frequency, and tends to zero as the π-mode cut-off frequency is approached. Hence the coupling impedance becomes very large in the neighbourhood of the π-mode cut-off frequency, but this is achieved at the expense of the tuning rate. Similar remarks can be made about the T-ladder line, in which the electrons are coupled to the reverse space harmonic with θ_t between 0 and π.

In the interdigital line the electrons can be coupled to the backward space harmonic with θ_i between 0 and π. The line can be usefully compared with a monofilar helix having the dimensions $A_m = 2A_i$, $D_m = 2D_i$. For the same beam voltage, the frequency is roughly the same in both cases and $\theta_i \simeq \theta_m/2$ [see Fig. 4(*d*)]. It follows from eqn. (25) that $F_i(\theta_i) \simeq F_m(\theta_m/2)$. Hence, from Fig. 3, the f.d.f. in the line is generally very much greater than that in the helix. The group velocity has the same order of magnitude in both cases. It must be remembered, however, that $A_i = A_m/2$; hence, for the same current density, the total current which can be fed into the region of high field strength in the line is only half that which can be fed into the region of high field strength in the helix. Taking this into account, we see that, for the same frequency, the figure of merit $AK(\theta)$ is roughly the same as that in a bifilar helix with $A_b = 2A_i$ and $D_b = D_i$.

In the meander line the electrons can be synchronized with the backward space harmonic with θ_s between π and 2π. From the point of view of backward-wave operation, the meander line is roughly equivalent to a monofilar helix with the same dimensions.

There are many factors to be taken into account in the design of a travelling-wave tube. A much more detailed discussion than has been given here would be necessary before choosing the tape structure appropriate to any particular tube. One can, however, draw a few general conclusions. There are no serious rivals to the monofilar helix for forward-wave operation, except with regard to their mechanical and thermal properties. The bifilar helix has a larger forward-wave coupling impedance at high frequencies, but backward-wave oscillations may occur due to coupling to the push-pull mode. The T-ladder line is very narrow-banded in forward-wave operation, but it is more rugged than a helix and has a higher thermal capacity. For backward-wave operation the interdigital line is an excellent structure. Its good mechanical and thermal properties make it superior to the bifilar helix. Ladder lines also have good mechanical and thermal properties, but they have lower tuning rates than the interdigital line. The meander line is probably only of academic interest.

(8) CONCLUSION

It has been shown that the exact solution to the problem of TEM wave propagation along a periodic array of parallel straight tapes can be applied easily to the determination of the coupling impedances and dispersion curves of a variety of tape structures. More refined analyses of these structures than have been made here could be based on the gap fields (or the fields over the tapes) which we have found. The treatment of undeveloped helices due to Sensiper,[9] for example, starts out from an assumed field distribution in the gaps (or over the tapes). For the purpose of discussion it is convenient to concentrate attention on the gap containing the origin of co-ordinates. Sensiper assumes that the component of the electric field parallel to the edge of the gap is zero, and makes various assumptions about the component of the electric field normal to the edge of the gap. In the most sophisticated treatment, the amplitude of this component is taken to have the form $(z^2 - s^2/4)^{-1/2}$ so as to introduce the appropriate singularities at the edges of the tapes. This is not very

different from the amplitude factor in eqn. (14), particularly when the gap width is small. The phase factor is either taken to be a constant, i.e. there is no change of phase across the gap, or it is taken to have the form $\exp(-j\theta z/D)$, which differs from the phase factor in eqn. (14) in that θ replaces $\theta - \pi$. It would be interesting to see if Sensiper's results are changed very much when this inaccuracy is removed from his treatment. The work of Tien[3,11] could be subjected to the same revision. Watkins and Ash[6] consider developed helices and use the constant-field approximation (see Section 11.3). Developed helices have been treated exactly in Section 4 of the present paper.

When θ lies between $1 \cdot 3\pi/2$ and 2π, the exact value of the f.d.f. of a tape structure is considerably greater than the approximate value obtained on the basis of the constant-field approximation. This revision will have a marked effect on the theoretical performance predicted for tubes in which θ lies in this region. For example, the theoretical starting current of a Q-band (8·6mm) backward-wave oscillator under development at the Radar Research Establishment[5] is about 4mA or 1mA, according as we use the constant-field approximation to the f.d.f. or the value found here, assuming 25dB distributed loss and $\theta = 7\pi/4$ in both cases.

(9) ACKNOWLEDGMENTS

The author would like to thank H. W. Duckworth and J. S. Thorp, of the Radar Research Establishment, and J. H. Collins and F. Gill, of The General Electric Co., Ltd., for many stimulating discussions.

(10) REFERENCES

(1) PIERCE, J. R.: 'Traveling-Wave Tubes' (D. Van Nostrand, New York, 1950).

(2) CHRISTENSEN, W. V., and WATKINS, D. A.: 'Helix Millimeter-Wave Tube', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 93.

(3) TIEN, P. K.: 'Bifilar Helix for Backward-Wave Oscillators', *ibid.*, 1954, **42**, p. 1137.

(4) KARP, A.: 'Traveling-Wave Tube Experiments at Millimeter Wavelengths with a New, Easily Built Space Harmonic Circuit', *ibid.*, 1955, **43**, p. 41.

(5) DUCKWORTH, H. W., and THORP, J. S.: Radar Research Establishment, unpublished.

(6) WATKINS, D. A., and ASH, E. A.: 'The Helix as a Backward-Wave Circuit Structure', *Journal of Applied Physics*, 1954, **25**, p. 752.

(7) FLETCHER, R. C.: 'A Broad-Band Interdigital Circuit for use in Traveling-Wave-Type Amplifiers', *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 951.

(8) BUTCHER, P. N.: 'A Theoretical Study of Propagation Along Tape Ladder Lines' (see page 169).

(9) SENSIPER, S.: 'Electromagnetic Wave Propagation on Helical Structures', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 149.

(10) LEBLOND, A., and MOURIER, G.: 'Étude des lignes à structure périodique', *Annales de Radioélectricité*, 1954, **9**, p. 180, Part 2, p. 311.

(11) TIEN, P. K.: 'Travelling-Wave Tube Helix Impedance', *Proceedings of the Institute of Radio Engineers*, 1953, **41**, p. 1617.

(12) MONTGOMERY, C. G., *et al.*: 'Principles of Microwave Circuits' (McGraw-Hill, 1948), p. 28.

(13) JAHNKE, E., and EMDE, F.: 'Tables of Functions' (Dover Publications, 1943).

(14) PIERCE, J. R.: 'Propagation in Linear Arrays of Parallel Wires', *Transactions of the Institute of Radio Engineers*, 1955, **ED|2**, p. 13.

(11) APPENDICES

(11.1) Application of Babinet's Principle

The tape array shown in Fig. 1 has gap width s and tape width w. It can propagate a TEM travelling wave, with phase change θ per period of the array, for which the electric and magnetic potential functions on the plane $x = 0$ are U and V, respectively, as calculated in Section 2. Consider now the 'complementary' array which is derived from that shown in Fig. 1 by interchanging the gaps and the tapes. The complementary array has gap width w and tape width s. Moreover, inspection of eqns. (6) and the boundary conditions (7) shows that it can propagate a TEM travelling wave, with phase change θ per period of the array, for which the electric and magnetic potential functions on the plane $x = 0$ are, respectively,

$$\left.\begin{aligned} U' &= V \\ V' &= -U \end{aligned}\right\} y > 0 \quad . \quad . \quad . \quad . \quad . \quad (36)$$

This is an example of Babinet's principle for electromagnetic fields.[12] The similarity of the formulae for $E_z(0, 0, z)$ and $H_z(0, 0, z)$—eqns. (14) and (17)—and the occurrence of $U(w/D, \theta)$ in eqn. (20*b*) are immediate consequences of eqns. (36).

It was remarked after eqn. (25) that $F(\theta)$ is unchanged when the gaps and the tapes are interchanged. This is also a consequence of Babinet's principle. In virtue of the boundary condition (7*a*), U and V can be expanded as series of space harmonics which have the forms

$$\left.\begin{aligned} U &= \sum_n u_n \exp -(j\beta_n z + |\beta_n| y) \\ V &= \sum_n v_n \exp -(j\beta_n z + |\beta_n| y) \end{aligned}\right\} y > 0 \quad . \quad . \quad (37)$$

because both functions satisfy Laplace's equation and tend to zero when y tends to infinity. Here, n takes all integral values,

$$\beta_n = (\theta + 2n\pi)/D \quad . \quad . \quad . \quad . \quad . \quad (38)$$

is the phase-change coefficient of the nth space harmonic, and u_n and v_n are coefficients which determine the amplitude and phase of the nth space harmonic. Substituting the expansions (37) into eqns. (6), and comparing coefficients, it is found that

$$v_n = j\frac{\beta_n}{|\beta_n|}u_n \quad . \quad . \quad . \quad . \quad . \quad . \quad (39a)$$

and so

$$|v_n| = |u_n| \quad . \quad . \quad . \quad . \quad . \quad . \quad (39b)$$

It follows immediately that the amplitudes of the various space harmonics are not altered by the transformation (36). Consequently $F(\theta)$ is not altered by this transformation.

Another interesting result can be derived in the case when the gap width is equal to the tape width. Then, interchanging the gaps and the tapes simply moves the array along the z-direction through half a period. Hence $U'(y, z)$ can differ from $U(y, z + D/2)$ only by a constant factor, μ say. Thus, from eqn. (36),

$$V(y, z) = \mu U(y, z + D/2) \quad . \quad . \quad . \quad . \quad (40)$$

Substituting the expansions (37) into this equation, comparing coefficients and using eqns. (38) and (39*a*), we find that

$$j\frac{\beta_n}{|\beta_n|}u_n = \mu u_n(-1)^n \exp(-j\theta/2) \quad . \quad . \quad . \quad (41)$$

Now, from eqn. (23), u_0 is not zero. Hence, putting $n = 0$ in eqn. (41) gives

$$\mu = j \exp(j\theta/2) \quad . \quad . \quad . \quad . \quad . \quad (42)$$

so that eqn. (41) reduces to

$$u_n\left[\frac{\beta_n}{|\beta_n|}-(-1)^n\right]=0 \quad . \quad . \quad . \quad . \quad . \quad (43)$$

when $n \neq 0$. It follows that

$$\left.\begin{array}{l} u_n = 0, n = 1, 3, 5 \ldots \infty \\ \quad -2, -4, -6 \ldots \infty \end{array}\right\} \quad . \quad . \quad . \quad . \quad (44)$$

i.e. all the odd positive harmonics and all the even negative harmonics are absent from the field expansions when the gap width is equal to the tape width.

A meander line can be derived from a 'complementary' interdigital line by interchanging the gaps and the tapes, replacing electric short-circuits by magnetic ones, and vice versa. Babinet's principle can be used to derive the field distribution in the meander line from that in the interdigital line. It is then a simple matter to verify the results which are stated at the end of Section 8.

(11.2) Evaluation of $\Delta U(s/D, \theta)$

From eqn. (19*b*)

$$U(s/D, \theta) = 2\int_0^\alpha dt \frac{\cos(1-\theta/\pi)t}{\sqrt{(\sin^2\alpha - \sin^2 t)}} \quad . \quad . \quad . \quad (45)$$

where
$$\alpha = \frac{\pi s}{2D} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (46)$$

As far as the author is aware, this integral cannot generally be expressed by a closed formula in terms of well tabulated functions. In the particular cases $\theta = 0$, $\pi/2$ and π, however, it is found, after some manipulation, that

$$U(s/D, 0) = \pi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (47)$$

$$U(s/D, \pi/2) = \frac{2K(\tan\alpha/2)}{\cos\alpha/2} \quad . \quad . \quad . \quad . \quad . \quad (48)$$

$$U(s/D, \pi) = 2K(\sin\alpha) \quad . \quad . \quad . \quad . \quad . \quad (49)$$

Here $K(\kappa)$ is the complete elliptic integral of the first kind[13] with modulus κ; it is not to be confused with the coupling impedance $K(\theta)$. For other values of θ an extremely good approximation can be derived. Integrating $U(s/D, \theta)$ by parts gives

$$U(s/D, \theta) = 2\cos[(1-\theta/\pi)\alpha]K(\sin\alpha)$$
$$+ 2(1-\theta/\pi)\int_0^\alpha dt \sin(1-\theta/\pi)t \int_0^t \frac{d\tau}{\sqrt{(\sin^2\alpha - \sin^2\tau)}} \quad . \quad . \quad . \quad . \quad . \quad (50)$$

The second term in this equation vanishes when $\theta = \pi$ and it is always considerably less than the first term. Moreover, any error made in approximating to the integrand in this term will be smoothed out by the double integration. It is convenient to make the approximation

$$(\sin^2\alpha - \sin^2\tau)^{-1/2} \simeq L(\alpha)(\alpha^2 - \tau^2)^{-1/2} \quad . \quad . \quad (51)$$

where $L(\alpha)$ is a function of α which will be chosen in a moment. Reversing the integration by parts *after* making this approximation gives

$$U(s/D, \theta) = 2\cos[(1-\theta/\pi)\alpha]K(\sin\alpha)$$
$$+ \pi L(\alpha)\{J_0[(1-\theta/\pi)\alpha] - \cos[(1-\theta/\pi)\alpha]\} \quad . \quad (52)$$

where J_0 denotes the zero-th order Bessel function. The quantity $L(\alpha)$ may be chosen so that this formula gives the correct answer when $\theta = 0$. Thus, from eqns. (47) and (52),

$$L(\alpha) = \frac{1 - \frac{2}{\pi}\cos\alpha K(\sin\alpha)}{J_0(\alpha) - \cos\alpha} \quad . \quad . \quad . \quad (53)$$

It is generally difficult to estimate the error introduced by the approximation (51). It is zero when θ is equal to 0 or π, and when $\theta = \pi/2$ we find numerically, from eqns. (48) and (52), that the error is much less than 1%. Direct numerical evaluation of the integral $\Delta U(s/D, \theta)$ confirms that the error is negligible for all the values of θ and α considered in this paper.

(11.3) The Constant-Field Approximation

The field of a TEM wave is symmetrically distributed about the plane of the array, $y = 0$. Hence, the mean stored energy in length A in the x-direction per unit length in the z-direction is

$$W = \frac{2}{D}\int_0^A dx \int_0^\infty dy \int_0^D dz \tfrac{1}{4}[\epsilon_0(|E_x|^2 + |E_y|^2) + \mu_0(|H_x|^2 + |H_y|^2)] \quad . \quad . \quad . \quad . \quad . \quad (54)$$

where ϵ_0 and μ_0 are, respectively, the permittivity and permeability of free space. Using eqns. (4), (5), (37) and (39), and carrying out the necessary integrations, we find

$$W = \epsilon_0 A \sum_{n=-\infty}^{\infty} |u_n|^2|\beta_n| \quad . \quad . \quad . \quad . \quad . \quad (55)$$

Moreover, calculating $E_z(0, 0, z)$ from eqns. (4), (5) and (37) and substituting into eqn. (22), we find that

$$E(\theta) = j(\theta/D)u_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (56)$$

on the line $x = 0$, $y = 0$. Finally, substituting from eqns. (55) and (56) into eqn. (3) gives

$$F(\theta) = |u_0|^2(2A\sum_{n=-\infty}^{\infty}|u_n|^2|\beta_n|)^{-1} \quad . \quad . \quad . \quad (57)$$

This equation is exact but u_n remains to be calculated. The constant-field approximation rests on the assumption that E_z is independent of z within each slot. Thus, if E_z is set equal to the convenient value D/s on the line $x = 0$ in the slot containing the origin of co-ordinates, it can easily be shown from eqns. (4), (5) and (37) that

$$u_n = \frac{1}{j\beta_n}\left(\frac{\sin\beta_n s/2}{\beta_n s/2}\right) \quad . \quad . \quad . \quad . \quad . \quad (58)$$

Substituting this in eqn. (57) gives the constant-field approximation to the f.d.f. at the array for a TEM travelling wave. Notice that, when θ tends to zero, $\beta_n|u_n|^2$ tends to infinity when $n = 0$, and remains finite when $n \neq 0$. Hence, when θ is small, only the zero-th space harmonic contributes to the stored energy, and eqns. (57) reduces to

$$F(\theta) \sim D/(2A\theta), \theta \to 0 \quad . \quad . \quad . \quad . \quad . \quad (59)$$

It will be clear that this result is rigorously correct; it does not depend on the constant-field approximation.

(11.4) The Ridge Factor

The f.d.f. of a ridge or T-ladder line is equal to that of an easitron structure multiplied by the ratio r of the stored energy per unit length in the easitron structure to that in the modified

structure; A, s, w, θ and $E(\theta)$ being made the same in both cases. A rough calculation of this ratio can be carried out on the basis of the simple expressions for the field vectors and the corresponding dispersion equations which rise in an analysis similar to Pierce's.[14] The results are particularly simple when the total channel width is equal to half the length of the tapes. Then, in the notation of the previous paper,[8] for the double-ridge structure,

$$r = \tanh \beta B \quad . \quad . \quad . \quad . \quad . \quad (60a)$$

For the single-ridge structure,

$$r = 2/(\coth \beta B + \coth \beta h) \quad . \quad . \quad . \quad (60b)$$

For the double-T structure,

$$r = \tanh \beta(B + b) \quad . \quad . \quad . \quad . \quad . \quad (60c)$$

And for the single-T structure,

$$r = 2/[\coth \beta(B + b) + \coth \beta h] \quad . \quad . \quad (60d)$$

Thus the ridge factor r will be appreciably less than one in ridge structures with practical dimensions because B is then relatively small, but it will be very close to one in T-structures provided that b and h are relatively large. The approximate formulae (60) apply to the centre line of the array. However, the error entailed by applying them elsewhere on the array is not likely to exceed that inherent in the formulae themselves.

DISCUSSION ON 'SERVO-OPERATED RECORDING INSTRUMENTS'*

Dr. R. L. Gordon (*communicated*): There appears to be an important omission from Dr. Maddock's paper. While treating the co-ordinate-plotting, or x–y, recorders in some detail, he makes no reference to the ability of the self-balancing potentiometer-recorder to record ratios of two signals directly, and at least one manufacturer markets an instrument intended specifically for such applications. It is clear that, since a self-balancing potentiometer indicates the ratio of an input signal to some constant fraction of a signal supplied by an internal cell (B_1 in Fig. 9A of the paper), it may equally well record the ratio of two input signals, if the 'denominator' signal replaces the internal cell B_2. Appropriate 'scaling' and 'true-zero' resistors (but see below) are required, in a similar manner to those of Fig. 9A: and arrangements are necessary for continuous modification of the gain round the feedback loop according to the instantaneous value of the denominator signal if this may vary considerably. In the commercial instrument referred to above, gain control is achieved by means of a magnetic amplifier. Although the use of 'true-zero' resistors requires that at least one of the input signals shall be 'floating', a method has been developed† in which the displacement of the zero point a short distance along the slide-wire is achieved by mixing with the numerator signal a small fraction of the denominator signal; where the amplitudes and impedances of the input signals permit this course, both signals may then have a common 'earth'—a useful feature where both are derived from valve circuits. The recording of ratios in this manner is of considerable value in optical instruments of the 'double-beam' type, in which the ratio of the light intensities transmitted along two different paths is usually required; a typical example is the double-beam infra-red spectrometer.

Dr. A. J. Maddock (*in reply*): The instrument mentioned by Dr. Gordon is, basically, a normal self-balancing potentiometer such as is described in Section 4 of the paper, but with the steady slide-wire current replaced by a current derived from the 'denominator' signal. Modification of connections to enable the instrument to be used in this way as a ratiometer can be made in instrument C2 of Table 2, but it is not a facility available generally in other makes.

In the function, or x–y, recorders, both signals can be reduced to zero, but this is not the case in the simpler ratiometer mentioned above. The 'numerator' signal can be reduced to zero, but the 'denominator' signal can only be reduced so far, otherwise the ratio of the signals would tend to infinity: in practice a possible reduction to about 1/40 of its maximum value is adopted.

* MADDOCK, A. J.: Paper No. 2131, September, 1956 (see 103 B, p. 617).
† GORDON, R. L.: 'The Use of Self-Balancing Pen Recorders as Ratiometers', *Journal of Scientific Instruments*, 1953, 30, p. 431.

621.317.742.029.64

Paper No. 2312 R
Mar. 1957

A COAXIAL STANDING-WAVE DETECTOR FOR THE S-BAND

By L. W. SHAWE, B.Sc., and G. W. FYNN.

(*The paper was first received* 10*th August, and in revised form* 22*nd November*, 1956.)

SUMMARY

A description is given of a coaxial standing-wave detector designed primarily for routine testing of coaxial components over the waveband 7·9–11 cm. It is terminated in a plug and socket of Pattern 6, RCL322, and is capable of measuring voltage standing-wave ratios up to 0·98. The design has been kept as simple as possible in order to make the instrument relatively easy to manufacture.

(1) INTRODUCTION

During the development of S-band test equipment the need arose for a coaxial standing-wave detector capable of measuring mismatches of the order of 0·98 v.s.w.r. The instrument was required to be fairly simple to make and to be suitable for production testing of coaxial components ending in plugs or sockets of Pattern 6, RCL322.[1] Since coaxial cable has a match no better than about 0·90 v.s.w.r., an accuracy higher than that quoted above would clearly be superfluous.

(2) DESCRIPTION

(2.1) General Description

The principles of design of standing-wave detectors have been dealt with by many authors during the last few years.[2–5] In general, all standing-wave detectors should be capable of easy insertion into, or connection to, the system whose characteristics it is desired to measure and then be capable of making measurements of standing-wave ratio to the stipulated accuracy.

The first requirement is easily met by ending the slotted section in a Pattern 6 plug at one end and a socket at the other. Then, by simply reversing the section it is possible to provide a plug or a socket ending as desired. The second requirement necessitates the insertion into the line of a probe which can travel parallel to the axis of the line and protrude a constant distance into it. Furthermore the conductors of the line must be accurately coaxial and the slotted section itself must be as free from discontinuities as possible. These somewhat contradictory requirements are not easy to meet.

The slotted section itself can be made of either an air-filled or a solid dielectric line. If the former is used, broad-band matching sections are required to match from the solid dielectric of the plug and socket to the air, and all the well-known difficulties of an air-spaced coaxial standing-wave detector would have to be surmounted. However, if the latter form is used these difficulties are largely overcome.

For easy self-checking of the instrument as regards wear, etc., the slotted section should be more than three half-wavelengths long at the longest wavelength at which it is intended to use it, namely 11 cm. The use of solid dielectric instead of air in the slotted section considerably shortens the length needed to meet this requirement.

For simplicity and the avoidance of possible sources of mismatch, the instrument (Fig. 1) was made with the inside diameter of the outer and the outside diameter of the inner the same as those appertaining to the respective portions of the mating faces of the Pattern 6 plugs and sockets. Then for the impedance to remain the same, namely 71 ohms, the dielectric used had to have a permittivity the same as that of the dielectric used in the plugs and sockets. This substantially reduced the choice of material, and of the few alternatives polythene was chosen as being the most readily available at the time. If a material of different permittivity had been used for the slotted section itself, matching sections and changes in diameter of inner or outer would have been needed, and the essential simplicity of the design would thereby have been lost.

Again in the interests of simplicity, the crystal-probe carriage was designed to use the outside of the coaxial line as a bearing surface, the spline giving location by virtue of its fit into the slot, rather than using more complicated and refined methods of carriage traversal and location. The slot was kept as narrow and shallow as possible in order to minimize the discontinuity effects. The crystal detector itself was chosen for its straightforward design and ready availability.

(2.2) Slotted Line

The outer of the slotted coaxial line (Fig. 2) was electroformed on a mandrel with a key of plastic material in it, the latter giving the slot on removal. By turning the outside to its final dimension before removing the mandrel, concentricity of the two walls was easily obtained.

The plug and socket outers were either made to be soldered on the finished tube or preferably made with tapers and then placed on the mandrel before plating and electroformed in position.

The total length of the slotted section is about 9 in, and to bore a hole in polythene just over $\frac{1}{16}$ in in diameter and 9 in long, to a positional accuracy of a mil, is not practicable, if even possible. The alternatives left are either to mould the polythene into position or to machine it in short lengths. The former method was rejected, owing to the difficulty of holding the inner accurately concentric when forcing the polythene into the mould, and also because of the danger of getting air bubbles trapped when moulding. The latter method has the big advantages of using rod polythene, thereby avoiding trouble from air bubbles, and of making it easy to achieve a high order of accuracy in the concentricity of the hole in the middle relative to the outside.

The turned sections of polythene, made about $\frac{1}{4}\lambda$ long at mid-band (i.e. 9·5 cm) so as to cancel out mismatches, are threaded on the inner, the surfaces and the faces between the sections being covered with Marco resin; the whole unit then inserted into the slotted tube, the Marco resin setting in position after assembly. By this means, any possible gaps between the polythene sections are eliminated and the sections are effectively keyed together, so that individual rotation is impossible. Moreover, the whole of the fabricated rod of polythene is to some extent stuck to the outer tube by virtue of the resin adhering to the metal and keying to the polythene.

The slot in the polythene is then milled, using the slot in the tube as a guide. The slot in the tube is $\frac{1}{16}$ in wide, so that the slot in the polythene is slightly less than this. The depth of the slot into the polythene is $\frac{1}{16}$ in, this being sufficient to ensure adequate coupling of the probe.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Shawe and Mr. Fynn are at the Radar Research Establishment.

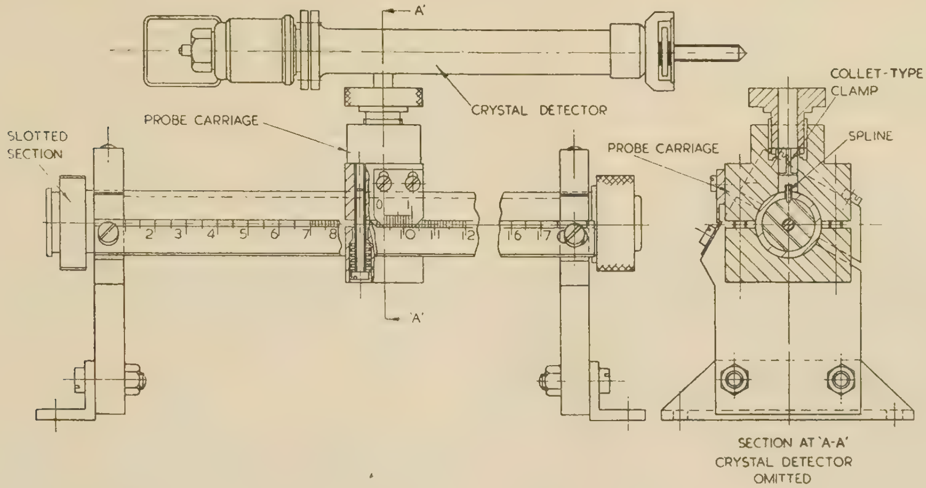


Fig. 1.—Partly-sectioned view of the detector.
The crystal detector is omitted from the section at AA.

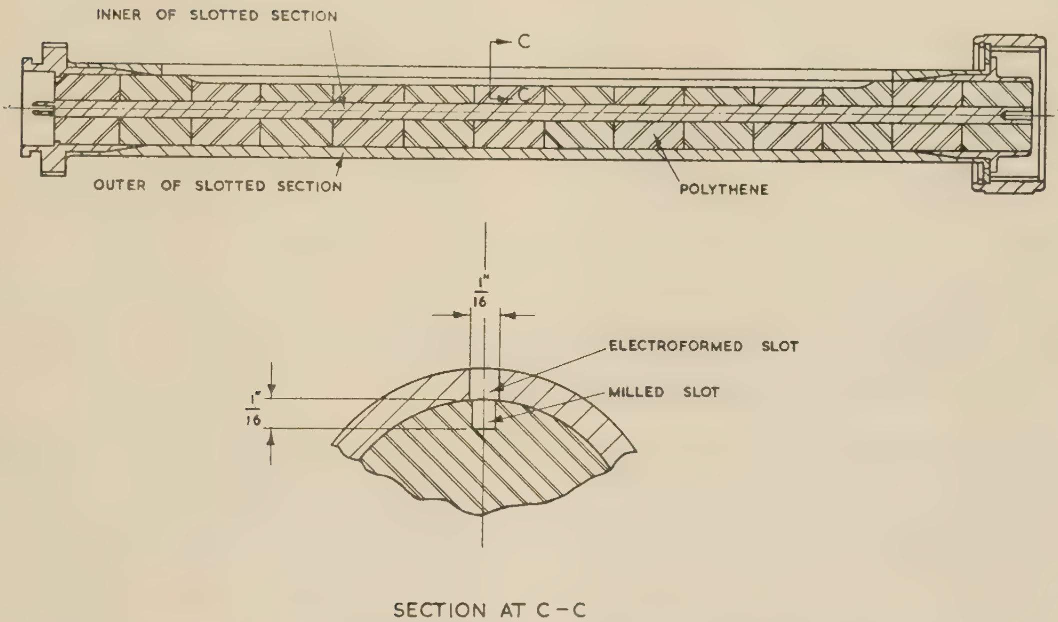


Fig. 2.—Sectional view of slotted line, with enlarged view across slot.

The inner of the coaxial line is made from a length of beryllium-pper, thus providing adequate springing for the inner contacts the plug and the socket, which are machined on the ends.

(2.3) Probe Carriage

The probe carriage is formed from a split block bored to a dius slightly greater than that of the outside of the slotted line, d carries a spline which fits into the slot. This spline must a good fit laterally, so as to prevent rock of the carriage on slotted line. The two halves of the block are spring loaded ther. A vernier carried on the block and working on to a nple fixed scale gives positional measurement of the probe ative to the end of the slotted section. The crystal detector is ld by a collet-type clamp in the top half of the block. The obe protrudes into the slot in the polythene through a small le in the spline, which is sleeved with Distrene for insulating purposes. The use of a collet-type clamp enables adjustment of probe penetration to be easily made with the simplest forms of crystal detector.

(2.4) Crystal Detector

The design of the crystal detector is quite conventional,[6] and is of simple form. The search probe is a wire 15 mils in diameter joined to a larger-diameter wire forming the inner conductor of a short length of coaxial line, which is terminated by the detector crystal and is tuned by a coaxial stub with a movable plunger.

(2.5) Finishes

The slotted coaxial line itself was given a rhodium flash on its outside, in order to protect it from wear and corrosion. The probe carriage was nickel plated, but the bore and spline were left in plain brass to give better wearing properties against the

rhodium. A life test on this combination of brass and rhodium showed no measurable wear after 10000 traverses of the probe carriage to and fro along the slotted line.

The inside surface of the slotted line and the inner conductor were left in their respective base metals.

(3) PERFORMANCE

Standing-wave detectors made on the above principles have given good performance results over the band 7·9–11 cm. When working into a given load, the total variation of the maxima or minima along the slot length will not exceed 0·25 dB, and with normal usage the worst error in voltage standing-wave ratio will not exceed 0·02 in volts.

The output of the crystal detector is fairly sensitive to variation of probe penetration. With a $\frac{1}{32}$in probe length, a withdrawal of 1 mil changes the crystal output by some 0·3 dB—hence the need for an accurate finish to both the probe carriage and the coaxial slotted section.

The output of the crystal detector is also varied by movement of the spline laterally in the slot. So long as this movement is kept less than 4 mils the output does not vary by more than 0·05 dB, but by working with the spline always against the same side of the slot considerably higher accuracy is naturally obtainable.

The impedance of the line is 71 ohms, and since the slot is narrow and shallow it remains sensibly of this value along the whole length of the slotted section.

Since the slot in the line is $6\frac{1}{2}$in (16·5 cm) long, it is possible to obtain three maxima or minima at 11 cm. This provides a convenient method of checking wear, etc., since all maxima and minima should be respectively equal—neglecting the attenuation of the line, which is less than 0·1 dB for the whole length.

For adequate output from the crystal detector, yet keeping the total crystal current less than 1 or 2 μA so that the crystal can always be regarded as giving square-law behaviour, the search probe protrudes about $\frac{1}{32}$in beyond the face of the masking spline.

The position of the probe along the slot can be measured to an accuracy of 0·01 cm.

The weakness of this design lies in the relative coefficients of thermal linear expansion of polythene and copper. There is a factor of about eight between the two materials, so that a quite small temperature change causes disastrous effects on the slotted section. For example, a rise of 20° C can cause the polythene to expand sufficiently to protrude beyond the end of the metal tube nearly a millimetre, but, owing to friction, when cooling again the sections tend first to stick and then to separate, producing a non-uniform impedance along the slotted line. Moreover, the walls of the slotted section are forced outwards, so that the slot widens and the diameter of the line increases, thereby permitting the spline excessive lateral movement in the slot, quite apart from a tendency for the carriage to ride on the wrong part of its bearing surface and hence to vary the penetration of the probe. However, these difficulties are not troublesome over a reasonable range of temperatures, and instruments made to the design have worked well and given no trouble from this cause when used in laboratories or factories in this country. It does, of course, prohibit their use by the Services, who require panclimatic design.

Although the design was primarily for use over the wavelength range 7·9–11 cm, the standing-wave detectors have been used satisfactorily down to 6 cm; naturally, tolerances on the slotted section—particularly the fit of the spline in the slot—become more critical, and greater care is needed in manufacture to ensure satisfactory results. The overall accuracy may also be slightly less. Its use at wavelengths above 11 cm is obviously possible, but the facility for easy checking (namely three maxima or minima) is then lost.

(4) CONCLUSIONS

The standing-wave detectors described have given satisfactory performance over the whole of the wavelength band 7·9–11 cm, and have met the requirements of being reasonably simple to make and capable of measuring mismatches up to 0·98 v.s.w.r.

They are capable of use on wavelengths greater than 11 cm, giving the same overall accuracy, but they then lose their ability of giving a built-in, and easy, method of check against wear, etc. They are also capable of use, to a slightly lower accuracy, on wavelengths considerably shorter than 7·9 cm.

The instruments are not suitable for tropical use, but they have given very good service, and performance, for over more than five years both in laboratory and factory when used on production testing.

(5) ACKNOWLEDGMENTS

Acknowledgment is made to the Chief Scientist, Ministry of Supply, and the Controller, H.M. Stationery Office, for permission to publish the paper.

The authors wish to express their thanks to Mr. T. Chaffe for the measurements which he has made on these instruments during the course of the development.

(6) REFERENCES

(1) 'Inter-Services Standards for Radio Components, List of Standard R.F. Plugs and Sockets, RCL322' (Radio Components Standardisation Committee, London, 1951).
(2) Clayton, R. J., Houldin, J. E., Lamont, H. R. L., and Willshaw, W. E.: 'Radio Measurements in Decimetre and Centimetre Wavebands', *Journal I.E.E.*, 1946, **93**, Part III, p. 107.
(3) Hunt, D., and Hogg, R. W.: 'The Design of Precision Standing-Wave Indicators for Measurements in Waveguides', *ibid.*, 1947, **94**, Part IIIA, p. 38.
(4) Barlow, H. M., and Cullen, A. L.: 'Microwave Measurements' (Constable, London), p. 118.
(5) Montgomery, C. G.: 'Technique of Microwave Measurements' (McGraw-Hill, New York), pp. 478 and 488.
(6) Reference 4, p. 134.

21.372.43:621.317.3.029.64

Paper No. 2310 R
Mar. 1957

AN S-BAND COAXIAL LOAD

By L. W. SHAWE, B.Sc.

(*The paper was first received* 10*th August, and in revised form* 22*nd November*, 1956.)

SUMMARY

The design and construction of a coaxial load terminated in a plug r socket of Pattern 6, RCL322 is described. The match gives a oltage standing-wave ratio better than 0·95 over the waveband -11 cm and can be improved with extra care during manufacture.

(1) INTRODUCTION

During the development of various coaxial items of test quipment for use in the waveband 7·9–11 cm, the need arose r a moderately good coaxial load giving a voltage standg-wave ratio better than 0·95. In particular, the load was equired to present a match which would be stable with time and e relatively robust, so as to overcome the disadvantages of oaxial cable loads. It was also required to be fairly easy and neap to make. All the instruments with which it was to be used nded in either a plug or socket of Pattern 6, RCL322.[1]

Coaxial loads using a cylindrical resistor have been described y Kohn[2] and Harris.[3] These loads have the advantage of orking over very large wavebands with a high order of match, ut require accurate and close-tolerance machining. That due Kohn would require a very small resistor (about 6 mm long) it were to work down to 6 cm wavelength, thus necessitating ry close tolerances. Loads using a tractrix contour for the ter (as suggested by Harris), although using a larger resistor, ve the disadvantages of relatively difficult machining, again to se tolerances, and also of requiring the resistor itself to be iformly distributed within narrow limits.

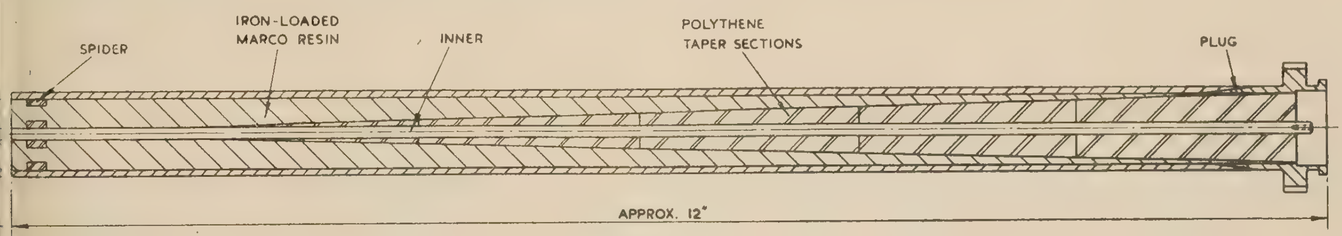
Since the load had to be used with Pattern 6 plugs and sockets, s logical it should be fitted with these types. With an aird line type of load such as those mentioned above, a further tching problem would therefore be involved. In order to n a reasonable overall match, that of the plug or socket to the air-spaced line would have to be of a fairly high order, thereby increasing the need for precise machining and hence the cost.

These considerations led to the development of the load described. Although not possessing the bandwidth of either of the resistor loads, it is easier to make, particularly in numbers, since a great deal of the precise machining is confined to the single operation of making a mandrel for electroforming the bodies. Furthermore, the finished load is very robust.

(2) DESCRIPTION

The design is shown in Fig. 1, which indicates the difference in ends for a plug and a socket.

The plug or socket outer is machined with a taper to a sharp edge at its back and with its inside diameter that of the mating face of the Pattern 6 range of plugs and sockets, namely 0·650 in. It is then placed on a stainless-steel mandrel which is slightly tapered (0·012 in in 12 in) and whose largest diameter is a little greater than 0·650 in. When the outer is pushed up against this taper it is held on the mandrel by friction. A copper tube is then electroformed so that the outer becomes integral with the tube. Any machining of the outside which may be required is done before the mandrel is removed.



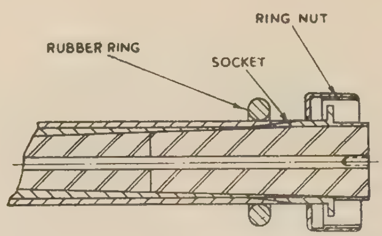


Fig. 1.—Coaxial load with plug and socket terminations.
The inset drawing shows the socket termination.

The inner is of the same diameter as the inner at the mating face of the Pattern 6 range, namely 0·110 in, and is made of a length of drawn beryllium-copper in the hard state so that the plug pin or socket receptacle can be machined directly on one end and will then have adequate springiness without further heat treatment.

Between the inner and the outer are five tapered sections of polythene, turned separately on a peg so that they are all individually concentric and, when assembled on the inner, will produce a smooth taper free from discontinuities. They are assembled with a film of Distrene cement, both between their faces and spread over the junctions, so that nothing can penetrate between the sections. The overall length of the taper is 10 in.

The taper assembly is placed in the tube with its large end,

Written contributions on papers published without being read at meetings are ed for consideration with a view to publication.
Mr Shawe is at the Radar Research Establishment.

which is slightly larger in diameter than the bore of the tube, making a force fit into the plug or socket. The polythene face is made flush with the appropriate shoulder of the plug or socket.

The other end of the inner is held concentric with the tube by a spider of polythene. The whole of the gap remaining in the tube is then filled with iron-loaded Marco resin which is allowed to harden. The mix used contains parts by weight: 100 of Marco resin 28, 5 of Monomer C, 4 of catalyst, 4 of accelerator and 300 of carbonyl-iron powder, grade M.E.

This method of construction enables a nearly perfect end-edge to the taper of attenuating material to be obtained at the point where lack of it would have most effect on the match, namely the input end of the load. The end of the polythene taper is removed, in terms of attenuation, sufficiently to have little effect on the match so that its sharpness is less critical. Thus the design is easy to manufacture. The avoidance of any air-filled section of coaxial line dispenses with the need for matching transformers, etc.

(3) PERFORMANCE

Loads to this design give a match better than 0·95 v.s.w.r. over the waveband 6–11 cm. The chief limiting factor to the match is the plug or socket, and if this is made to closer tolerances than normal, a better match can be obtained. Many loads have been made with a match better than 0·98 v.s.w.r. over the above waveband.

No measurements have been made above 11 cm, so that the ultimate upper wavelength for good matching is not known; nor have they been made below 6 cm, so the lower wavelength limit is also unknown. The impedance of the loads is 71 ohms, which is that of the Pattern 6 range of plugs and sockets.

(4) CONCLUSION

A simple, fairly easily made, coaxial load for use with instruments terminated in plugs or sockets of the Pattern 6 range has been described. It is capable of giving a match better than 0·95 v.s.w.r. over a waveband of at least 6–11 cm. With a little extra care this match may be increased to better than 0·98 v.s.w.r.

(5) ACKNOWLEDGMENT

Acknowledgment is made to the Chief Scientist, Ministry of Supply, and the Controller, H.M. Stationery Office, for permission to publish the paper.

(6) REFERENCES

(1) 'Inter-Services Standards for Radio Components, List of Standard R.F. Plugs and Sockets, RCL322' (Radio Components Standardisation Committee, London, 1951).

(2) Kohn, C. T.: 'The Design of a Radio-Frequency Coaxial Resistor', *Proceedings I.E.E.*, Monograph No. 83 R, November, 1953 (**101,** Part IV, p. 146).

(3) Harris, I. A.: 'The Theory and Design of Coaxial Resistor Mounts for the Frequency Band 0–4000 Mc/s', *ibid.*, Monograph No. 132 R, May, 1955 (**103 C**, p. 1).

21.372.543.2:121.372.413

Paper No. 2311 R
Mar. 1957

A RESONANT-CAVITY FILTER FOR THE S-BAND

By A. A. L. BROWNE, B.Sc.

(*The paper was first received* 10*th August, and in revised form* 22*nd November*, 1956.)

SUMMARY

A resonant-cavity filter has been developed which is tunable over e range 7·9–11·0 cm. It is terminated in an R.A.E. plug and socket ıd has a second-harmonic rejection of greater than 30 dB. The sertion loss is of the order of 0·2 dB, and the voltage standing-wave tio is better than 0·8 when measured in 71-ohm coaxial line.

(1) INTRODUCTION

In the measurement of transmitter powers an appreciable ror may be caused by the small amount of harmonics present the signal. This may, for example, be due to the use directional couplers, which frequently have much greater upling for the harmonics than for the fundamental. It is cessary, therefore, to provide a filter which will reject the inter-ring harmonics (mainly second) to such an extent as to render eir effect on the power measurement negligible. Such a filter uld be of considerable use in many laboratory measurements ıere the presence of harmonics is a common source of error.

(2) SPECIFICATION

The filter for this purpose was required to have an input ltage standing-wave ratio of better than 0·8 over the pass band 9–11·0 cm), and to have an attenuation of at least 30 dB to second harmonics of this range. The filter was to have as its terminals an R.A.E. plug and socket, and the requirements the v.s.w.r. were to apply to either terminal whilst the other s terminated in a 71-ohm coaxial load.

(3) DESIGN CONSIDERATIONS

It was found that a coaxial low-pass filter of reasonably short gth had too low an input voltage standing-wave ratio, and it s decided to use a tunable resonant-cavity filter with two coaxial uplings. The type of re-entrant cavity which consists of a cular cylinder tuned by means of a coaxial plunger (Fig. 1) was

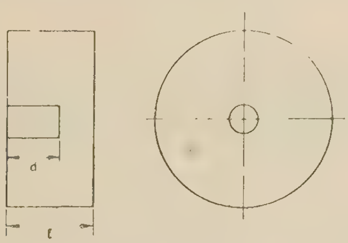


Fig. 1.—Form of re-entrant cavity used.

ught to be the most suitable for the purpose. Experimental k by Barrow and Mieher[1] has shown that on plotting resonant uency against d, the penetration of the plunger into the ity, there is obtained a series of gradual transitions between rfect cylinder ($d = 0$) and the perfect coaxial resonator where l (Fig. 2). Furthermore, an E-mode in the cylindrical state

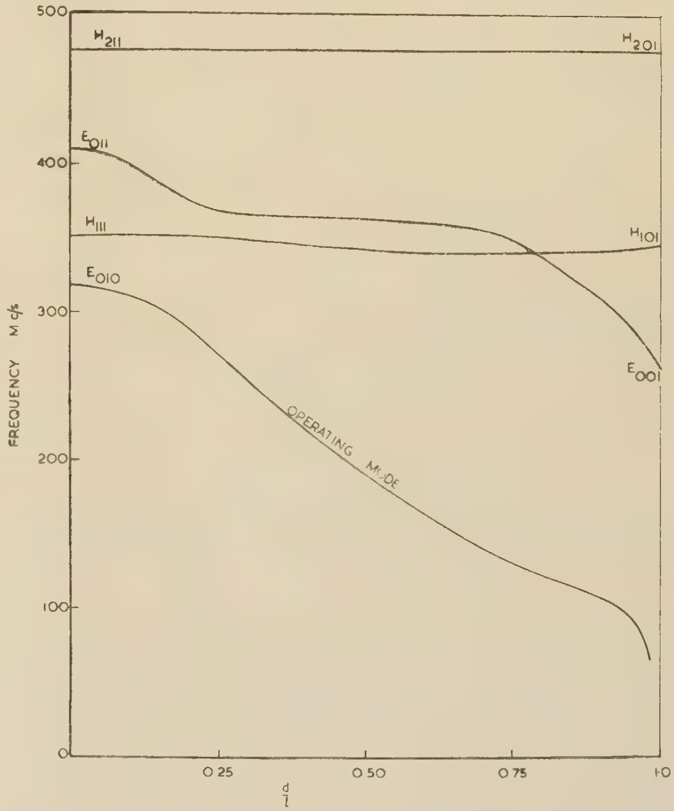


Fig. 2.—Some of the lowest modes of a re-entrant cavity. (After Barrow and Mieher.)

corresponds to an E-mode in the coaxial state, and an H-mode in the cylindrical state to an H-mode in the coaxial state. The E_{010}-mode in the perfect cylinder, however, gradually drops in frequency as the plunger is inserted until the frequency approaches zero as the plunger length d approaches the length l. In a suitably dimensioned cavity, therefore, it can be arranged that the resonant frequency of this mode over the tuning band is well separated from other resonances.

A cavity near to one resonance with two pairs of terminals can be shown to be equivalent[2] to the circuit of Fig. 3, where A and B are reference planes, determined by the particular cavity and coupling mechanism. The ratios n_1 and n_2 of the ideal transformers depend on the degree of coupling. In the case of symmetrical coupling, $n_1 = n_2$. For the particular cavity under consideration the input and output are connected to equal impedances and the coupling is symmetrical.

If therefore Q_0 and Q_L are the unloaded and loaded Q-factors of the cavity, and Z_0 is the impedance to which either arm is connected,

$$\frac{Q_0}{Q_L} = \frac{R + 2n^2 Z_0}{R} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

t itten contributions on papers published without being read at meetings are ited or consideration with a view to publication.
. Browne is at the Radar Research Establishment.

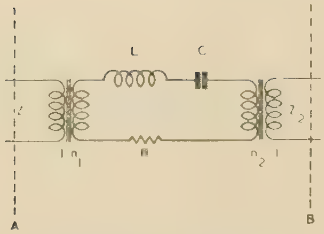


Fig. 3.—Equivalent circuit of coupled cavity.

For the v.s.w.r. at resonance,

$$\text{v.s.w.r.} = \frac{n^2 Z_0}{R + n^2 Z_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

From eqns. (1) and (2),

$$\text{v.s.w.r.} = \frac{Q_0 - Q_L}{Q_0 - Q_L} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The power loss, in decibels, in the cavity is

$$10 \log_{10} \frac{R + n^2 Z_0}{n^2 Z_0}$$

$$= 10 \log_{10} \frac{Q_0 + Q_L}{Q_0 - Q_L} \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Thus it is necessary to have a large ratio of unloaded to loaded Q in order to obtain a good match and a small insertion loss.

(4) METHOD OF MEASUREMENT

(4.1) Unloaded Q-Factor

The unloaded Q-factor of the cavity was measured by the Q-circle[3] method. The frequency was kept constant, and the variation of input impedance with cavity tuning positions was plotted on a Smith chart. The impedance was derived from standing-wave measurements in a No. 10 waveguide, the cavity being connected by means of a waveguide-to-coaxial transformer. The tuning plunger position was controlled by a 40 turns/in screw thread, small rotations of which were measured by having a small mirror attached and measuring the deflection of a spot of light. From these results the unloaded Q-factor of the cavity was derived.

There was some discrepancy between this method of measuring Q_0 and a frequency-sweeping method, in which the resonance curve was displayed on an oscilloscope and the bandwidth measured using an echo box as a high-Q wavemeter. The discrepancy was found to be due to too high a sweep frequency and agreement was obtained when this was reduced.

(4.2) Input V.S.W.R.

The v.s.w.r. of the cavity was measured using a coaxial standing-wave equipment and coaxial loads. Cable loads were found to be unsuitable and loaded Marco resin loads were used. To obtain symmetrical coupling, one probe was fixed in a particular position and the other adjusted until a similar v.s.w.r./wavelength graph was obtained using either probe as the input.

(4.3) Second-Harmonic Attenuation

To measure the attenuation of second harmonic through the cavity presented difficulty, since no oscillator was available on the particular frequency range. It was therefore decided to scale the cavity to double the size in all relevant details, and it was assumed that (apart from a small error due to the change in skin depth) the attenuation of this cavity to the range 7·9–11·0 cm was the same as that of the normal-size cavity to the second harmonics of this range. A rough graph of plunger position and tuning wavelength in the range 15·8–22 cm was plotted, and in the attenuation measurement at 7·9–11·0 cm the plunger was set in approximately the correct position for the corresponding wavelength. It was found that the attenuation of the cavity in the range 7·9–11·0 cm was relatively insensitive to change in the plunger position. The attenuation was measured by a substitution method using as a source an R.R.D.E. signal generator type 16.

(5) RESULTS

An original cavity of diameter 1·75 in and length 1 in tuned by a $\frac{3}{8}$ in plunger was found to have insufficient attenuation of second harmonic owing presumably to the higher modes of Fig. 2 being supported. It was reasoned that reducing the cavity length to considerably less than one-half of the shortest wavelength to be rejected would suppress all higher modes, except those E-modes with uniform distribution in the axial direction. To suppress these modes it was necessary that the diameter of the cavity should be sufficiently small. On reducing the cavity length to 0·5 in it was found that the second-harmonic attenuation was sufficient. Keeping the coupling symmetrical it was found that increasing the coupling gave a better v.s.w.r. performance over the band. The coupling was increased until the v.s.w.r. was satisfactory and the loaded Q-factor was still high enough to give sufficient second-harmonic rejection.

The cavity was made of brass, the inside being copper-plated in an acid copper bath. The final design is shown in Fig. 4. The original cavity was designed with a set of phosphor-

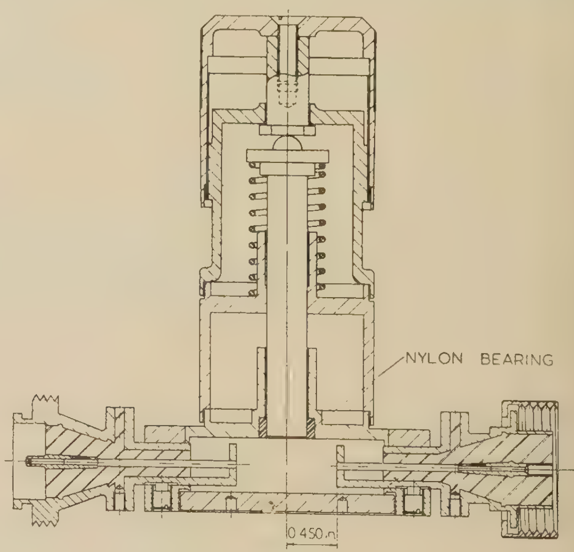


Fig. 4.—Final version of cavity.

bronze spring fingers to make contact with the plunger. The unloaded Q-factors of this type of cavity were measured and found to be rather low (500–800). By replacing the spring-finger arrangement by a choke and non-contacting plunger assembly it was found that the Q-factors were considerably improved. The value of Q_0 was measured at wavelengths of 8·0, 10·0 and 11·0 cm and found to be 2000, 2200 and 2100, respectively, with an estimated error of $\pm 5\%$. The input voltage standing wave ratio is shown in Fig. 5, and the second-harmonic attenuation is shown in Fig. 6 for the probe settings finally adopted.

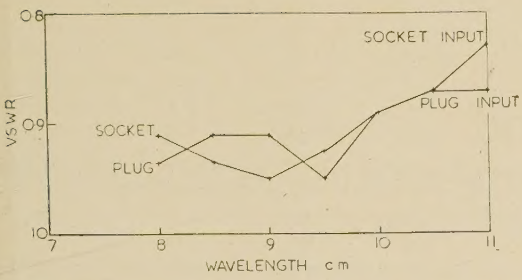


Fig. 5.—Input voltage standing-wave ratio of filter.

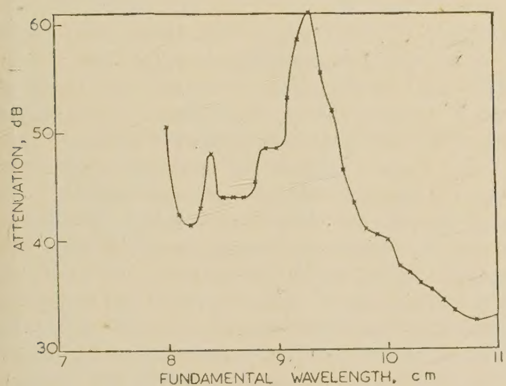


Fig. 6.—Attenuation of the second harmonic.

The loaded Q-factor was obtained by measuring the half-wer bandwidth and was found to vary from 32 to 42 over e band 7·9–11 cm.

At 10 cm the measured values of Q_L and Q_0 were 33 and 00.

Therefore,

$$\text{Power loss, in dB,} = 10 \log \frac{2233}{2167} \text{ (from eqn. 4)} = 0{\cdot}13$$

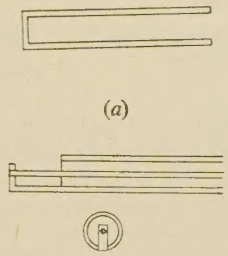
$$\text{Insertion loss} = 0{\cdot}13\,\text{dB} + \text{mismatch loss.}$$

can be seen that the inherent mismatch due to symmetrical upling [eqn. (2)] is masked by the effect of mismatches due to ugs and sockets.

The measured value of the insertion loss was found to be nstant over the band within the accuracy of measurement and be 0·16 ± 0·05 dB.

(6) MANUFACTURE

It was found that, in order that the coupling should remain mmetrical over the whole of the frequency band, the two upling loops should be made as nearly identical as possible. This was done by machining the loop and outer of the coaxial line from a piece of brass rod. The rod was drilled out as shown in Fig. 7(*a*), the square bottom being obtained by using a D-bit. The outer of the tube was then machined until the wall thickness



(*a*)

(*b*)

Fig. 7.—Method of manufacture of probes.

was correct. A small hole was then drilled in the centre of the end wall to take the inner of the coaxial line, and a portion of the tube wall was milled away leaving the loop of Fig. 7(*b*) (shown with the inner in place). This was found to be the most satisfactory way of making reproducible probes. Because manufacturing tolerances are not sufficient to enable the probes to be preset to the necessary degree of accuracy, the proposed method of setting up the cavity is to preset one of the probes and to set the other experimentally in such a position that the high v.s.w.r. obtained is the same at either end of the cavity at a wavelength of 9·5 cm (mid-band). It is considered necessary also to check the v.s.w.r. at 7·9 and 11·0 cm to ensure that it is still within the specification and preferably better than 0·83.

Some care is needed to ensure good contact between the end plate of the cavity and the walls. Fig. 4 shows the arrangement as used in the fully engineered version of the cavity.

(7) ACKNOWLEDGMENT

Acknowledgment is made to the Chief Scientist, the Ministry of Supply and to the Controller of H.M. Stationery Office for permission to publish the paper. The cavity was developed jointly by the R.R.E. and the Sperry Gyroscope Co., Ltd.

(8) REFERENCES

(1) BARROW, W. L., and MIEHER, W. W.: 'Natural Oscillations of Electrical Cavity Resonators', *Proceedings of the Institute of Radio Engineers*, 1940, **28**, p. 184.

(2) MONTGOMERY, DICKE and PURCELL: 'Principles of Microwave Circuits' (McGraw-Hill), M.I.T. series, **8**.

(3) ALTAR, W.: 'Q-Circles—A Means of Analysis of Resonant Microwave Systems I–II', *Proceedings of the Institute of Radio Engineers*, 1947, **35**, p. 355, p. 478.

DISCUSSION ON

'A DEVELOPMENT OF THE COLLARD PRINCIPLE OF ARTICULATION CALCULATION'*

Dr. J. Collard (*communicated*): As it is nearly thirty years since I carried out the work on which this paper is based, I am very interested to learn that my original theory still finds a use. I should like to restate an important principle which, I feel, had considerable influence in bringing the work to a satisfactory conclusion. It is that any constants introduced into formulae should correspond to actual physical quantities, and the form of any expression relating one quantity to another should be derived from theoretical considerations of the phenomena involved.

An example is the relation between word and sound articulation. I could have tried fitting some arbitrary mathematical function to the experimental results. Fortunately, however, I pondered over the problem of how a listener might be expected to recognize either random syllables or actual words and was led to the expression

$$w = \frac{1}{1 + k(1/a^n - 1)}$$

This expression is derived in the following way. Experience gained when taking part in articulation testing indicates that, when the circuit is not perfect, the listener is definitely in a position of having to make a choice from possible alternative sounds which seem equally probable to him. Thus, if a sound articulation of, say, $0 \cdot 5$ is obtained, this means that the listener gets a sound right on the average once in two times, and one can imagine him as considering the two possible alternatives and choosing one at random. In general, therefore, he will have to choose one out of $1/a$ possibilities. When the sounds are combined into syllables of n sounds, each sound will, on the average, have $1/a$ possible alternatives, so that there will be $(1/a)^n$ possible syllables from which one is chosen at random.

When the syllables used are known by the listener to be actual words, he is able to reject those alternatives that are not words and thus obtains a higher articulation. In this case, of the $1/a^n$ alternatives, one, the called syllable, must be a word, so there are $1/a^n - 1$ which may or may not be words. If, for all syllables of the type used, a fraction k are actual words, then the total number of actual words from which the listener has to choose one is $k(1/a^n - 1)$ plus the called word, so that the word articulation will be one out of $1 + k(1/a^n - 1)$. This expression agrees very closely with experimental results when k is determined for the type of syllable being used. It is reasonable to assume, therefore, that the above is a true description of what goes on in the listener's mind.

Having reached this point, it occurred to me that, just as words are made out of one or more sounds, it might be possible to consider sounds as being made out of a number of frequency components gathered together into one or more fairly isolated bands, which I called 'characteristic bands'. That this is so in the case of vowel sounds is obvious, and certain evidence exists that something of the sort occurs with consonants. This suggested to me that I could postulate a quantity, band articulation, which would be related to sound articulation just as sound articulation is related to word articulation. Band articulation is the average probability that a characteristic band will be correctly received over a given circuit. In this way, the problem of determining a subjective quantity, sound articulation, is reduced to the much simpler one of calculating the objective quantity, band articulation, which can thus be directly related to physical quantities such as frequency characteristic, noise level, etc.

The same considerations arise when applying the theory to the assessment of telephone transmission. In this case, the authors' curve of Fig. 3, which gives the proportion of band articulation as a function of sensation level, was, presumably, obtained from the results of a specially selected and trained crew, and is therefore too steep. To obtain a corresponding curve which would apply to actual subscriber conditions, I analysed the factors that affect this curve, such as the statistical variation of speaking level, acuity of hearing, distribution of power in speech sounds, etc., and obtained the much less steep curve which I reproduced in the authors' Reference 3.

Messrs. D. L. Richards and **R. B. Archbold** (*in reply*): We are very grateful to Dr. Collard for restating the principles upon which his theory is based. We cannot agree, however, that it is possible in practice to formulate a reasonably manageable framework purely from theory and to predict articulation results for actual crews without empirical determination of parameters. Some of the considerations that are difficult to include in any simple theory are as follows:

(*a*) The frequencies of occurrence of individual sounds are not uniform, so that taking an average value cannot be justified.
(*b*) Digram sound structure cannot be neglected.
(*c*) The probability of error varies widely from one sound to another.
(*d*) Errors of adjacent sounds are interdependent.
(*e*) When a sound is received incorrectly the alternatives are not equally probable.

Furthermore, it seems unlikely that band articulation would prove much less dependent upon a crew than are the directly measured sound or word articulations.

We prefer to deal with subscriber variations somewhat differently from Dr. Collard, namely to determine the proportion of calls which subscribers would find satisfactory, e.g. which achieve a given level of articulation, rather than the mean articulation averaged over all calls.

* Richards, D. L., and Archbold, R. B.: Paper No. 2143 R, September, 1956 (see 103 B, p. 679).

PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Part B. RADIO AND ELECTRONIC ENGINEERING (INCLUDING COMMUNICATION ENGINEERING) MARCH 1957

CONTENTS

PAGE

Very-Low-Frequency Propagation and Direction-Finding .... F. HORNER, M.Sc. 73
Scottish Centre: Chairman's Address .... PROF. F. M. BRUCE, M.Sc., Ph.D. 81
South Midland Centre: Chairman's Address .... C. J. O. GARRARD, M.Sc. 83
Acoustics of Large Orchestral Studios and Concert Halls .... T. SOMERVILLE, B.Sc., and C. L. S. GILFORD, M.Sc. 85
Frequency Diversity in the Reception of Selectively Fading Binary Frequency-Modulated Signals.
J. W. ALLNATT, B.Sc.(Eng.), E. D. J. JONES and H. B. LAW, B.Sc.Tech. 98
An Investigation of the Spectra of Binary Frequency-Modulated Signals with Various Build-up Waveforms.
J. W. ALLNATT, B.Sc.(Eng.), and E. D. J. JONES 111
An Improved Fading Machine .... H. B. LAW, B.Sc.Tech., F. J. LEE, R. C. LOOSER, B.Sc.(Eng.), and F. A. W. LEVETT 117
The Signal/Noise Performance Rating of Receivers for Long-Distance Synchronous Radiotelegraph Systems using Frequency Modulation.
H. B. LAW, B.Sc.Tech. 124
The Detectability of Fading Radiotelegraph Signals in Noise .... H. B. LAW, B.Sc.Tech. 130
Laboratory Test Equipment for Synchronous Regenerative Radiotelegraph Systems.
C. G. HILTON, H. B. LAW, B.Sc.Tech., F. J. LEE and F. A. W. LEVETT 141
Discussion on the above six papers 147
High-Temperature Properties of Tungsten which influence Filament Temperatures, Lives and Thermionic-Emission Densities.
R. N. BLOOMER 153
A Survey of Factors limiting the Performance of Magnetic Recording Systems.
E. D. DANIEL, M.A., P. E. AXON, O.B.E., M.Sc., Ph.D., and W. T. FROST 158
A Theoretical Study of Propagation along Tape Ladder Lines .... P. N. BUTCHER, Ph.D. 169
The Coupling Impedance of Tape Structures .... P. N. BUTCHER, Ph.D. 177
Discussion on 'Servo-Operated Recording Instruments' 187
A Coaxial Standing-Wave Detector for the S-Band .... L. W. SHAWE, B.Sc., and G. W. FYNN 188
An S-Band Coaxial Load .... L. W. SHAWE, B.Sc. 191
A Resonant-Cavity Filter for the S-Band .... A. A. L. BROWNE, B.Sc. 193
Discussion on 'A Development of the Collard Principle of Articulation Calculation' 196

Bibliographical References.—It is requested that bibliographical reference to an Institution paper should always include the serial number of the paper and the month and year of publication, which will be found at the top right-hand corner of the first page of the paper. This information should precede the reference to the Volume and Part.

Example.—SMITH, J.: "Reflections from the Ionosphere," *Proceedings I.E.E.*, Paper No. 3001 R, December, 1954 (102 B, p. 1234).

Published by The Institution, Savoy Place, London, W.C.2. *Telephone:* Temple Bar 7676. *Telegrams:* 'Voltampere, Phone, London.'
Printed by Unwin Brothers Limited, Woking and London.